UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

What's in a domain?

Towards fine-grained adaptation for machine translation

van der Wees, M.E.

Publication date
2017
Document Version
Final published version
License
Other

Link to publication

Citation for published version (APA):
van der Wees, M. E. (2017). *What's in a domain? Towards fine-grained adaptation for machine translation*. [Thesis, fully internal, Universiteit van Amsterdam].

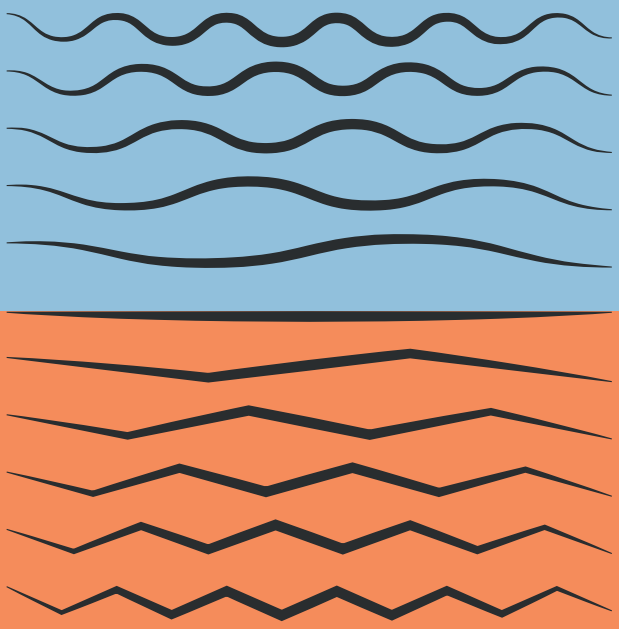
General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Nov 2026

WHAT'S IN A DOMAIN?
TOWARDS FINE-GRAINED ADAPTATION FOR MACHINE TRANSLATION



MARLIES VAN DER WEES

What's in a Domain?

Towards Fine-Grained Adaptation for Machine Translation

Marlies van der Wees

What's in a Domain?

Towards Fine-Grained Adaptation for Machine Translation

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Aula der Universiteit
op vrijdag 8 december 2017, te 13:00 uur

door

Maria Elisabeth van der Wees

geboren te Utrecht

Promotiecommissie

Promotor:		
	prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotores:		
	dr. C. Monz	Universiteit van Amsterdam
	dr. A. Bisazza	Universiteit van Amsterdam
Overige leden:		
	prof. dr. A.P.J. van den Bosch	Radboud Universiteit Nijmegen
	dr. R. Fernandez Rovira	Universiteit van Amsterdam
	dr. T.E.J. Mensink	Universiteit van Amsterdam
	prof. dr. K. Sima'an	Universiteit van Amsterdam
	prof. dr. J. Tiedemann	Universiteit van Helsinki

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Copyright © 2017 Marlies van der Wees, Amsterdam, The Netherlands
Cover by David Graus
Printed by Off Page, Amsterdam

ISBN: 978-94-6182-852-1

Acknowledgements

I did it! I successfully finished four years of research, the proof of which is in your hands right now. But like every PhD candidate, I didn't do it on my own, and there are a number of people whom I owe eternal gratitude. Some helped me with the actual research, while others were there mostly for mental support and the necessary distraction. Both have been incredibly important!

First, Christof thank you for giving me the opportunity to do this PhD. Coming from a different field, I failed to correctly answer many of your technical interview questions, but you still believed in me. Over the years, you not only taught me about research, but you also encouraged me to take a stance and fight for my ideas, helped me to better manage expectations, and supported me during the DIGIT meetings. I don't think many PhD students have eaten fish and chips with their advisor in an English pub.

Arianna, 'my postdoc,' thanks for getting on board on my projects. You were a truly amazing collaborator. Without your critical thinking and your pleasant company, my PhD would have been way less enjoyable. I am sure you are gonna have a great career in research and that I'll be bragging about the fact that I worked with the famous Prof. Dr. Bisazza in her early days.

The rest of the SMT group: Katya, Ke, Marzieh, Hamid, Praveen, and Ivan, thanks for sharing both fun and failure, having interesting discussions, proofreading each other's papers, and gossiping about our seniors.

Wouter, thanks for encouraging me to stop hesitating and just write that paper. It was exactly the support I needed. You showed me that it doesn't have to be that difficult to write a paper and, in fact, to finish a PhD.

Maarten, we never collaborated, but I want to thank you for running the great research group that I've been part of, and for providing valuable feedback on my presentations and thesis.

All other current and former ILPS folks, thanks for making my PhD experience complete. Both the social and the sporty events were additions that I wouldn't have wanted to miss. A few people have particularly improved my life as a PhD student: David, thanks for being my ILPS-bestie and sharing your home-baked banana bread with me during our ubiquitous tea breaks. Anne, Daan and Tom, thanks for being older and wiser, asking critical questions, and encouraging me to keep improving. Evgeny, thanks for your enthusiasm to join me in reaching my goal of running 1000 kilometers in a year.

I also want to thank Nigel, Rachel, Rob, Sue and others involved in the DIGIT project. Every meeting left me with countless ideas to improve my research, and helped me to keep the bigger picture in mind.

Antal, Jörg, Khalil, Rachel and Thomas, thanks for being my committee members, and for being among the very few people who actually read (most of) this thesis. Kitty and Marzieh, my paranymphs, thanks for willing to 'fight' this committee together. With two such strong women on my side, the defence can only be a piece of cake!

Speaking of strong women, I have quite a lot of those around me, and I'm going to mention some of them here: Ruth, mijn reismaatje. Onze trips naar Colombia en Indonesië waren precies wat ik nodig had, en je bent echt het beste reismaatje dat ik

me kan voorstellen. Laten we snel weer die tassen pakken en materiaal voor een nieuw fotoboek verzamelen!

De meiden van Lust: Denise, Kitty, Lujzika en Valentina. Onze wine-and-dine en boer-zoekt-vrouw avondjes zijn de perfecte entertainment en ontspanning in drukke tijden. Denise, bedankt voor de vele logeerpartijtjes toen we elkaar nodig hadden.

De Carré-meiden: Ellen, Puikay, Sanne en Sophie. Ik geniet al jaren van onze etentjes, weekendjes weg, barbecues op de tuin en jullie leuke kids, maar ben vooral ook intens dankbaar voor jullie steun in moeilijkere tijden. Laten we nog lang zo doorgaan met z'n vijven!

Sigrid, mijn bootcampmaatje. Lekker ravotten in het park is nog zoveel leuker in goed gezelschap. En al helemaal als we onszelf na afloop belonen met een IJwitje en zoete-aardappelfrieten in een van de vele hipstercafé's in Amsterdam-Oost.

Anna, bedankt dat je altijd bereid bent om Bamie te verzorgen, je hebt me een hoop stress ontnomen. Ik blijf je meenemen naar De Parade tot we oud en grijs zijn.

Ten slotte Papa, Mama en Bernard. Het zal niemand verbazen dat jullie altijd in me hebben geloofd en ook tijdens mijn PhD heb ik me erg gesterkt gevoeld door jullie steun en betrokkenheid. Jullie scheppen met trots op over jullie 'slimme dochter' (en zus), en bewaren krantenknipsels over alles wat met 'big data' of de UvA te maken heeft.

Dennis, bedankt dat je in mijn leven bent verschenen. Ik heb in korte tijd al zó veel van je geleerd dat ik bijna nog een proefschrift kan schrijven. Mede dankzij jou waren de laatste loodjes niet het zwaarst!

Marlies
October 2017

Contents

1 Introduction **1**
- 1.1 Research outline and questions . . . 2
- 1.2 Main contributions . . . 5
- 1.3 Thesis overview . . . 7
- 1.4 Origins . . . 9

2 Background **11**
- 2.1 Parallel and monolingual corpora . . . 11
- 2.2 Phrase-based machine translation . . . 12
  - 2.2.1 Word alignment and phrase extraction . . . 12
  - 2.2.2 Translation models . . . 13
  - 2.2.3 Reordering models . . . 14
  - 2.2.4 N-gram language models . . . 15
  - 2.2.5 Decoding and the log-linear model . . . 16
- 2.3 Neural machine translation . . . 18
  - 2.3.1 Word embeddings and vocabulary size . . . 18
  - 2.3.2 Recurrent neural language models . . . 19
  - 2.3.3 Encoder-decoder models . . . 19
  - 2.3.4 Attention mechanisms . . . 20
  - 2.3.5 Training an NMT system . . . 22
  - 2.3.6 Decoding . . . 23
- 2.4 Experimental setup . . . 24
  - 2.4.1 Oister and general PBMT setup . . . 24
  - 2.4.2 Tardis and general NMT setup . . . 24
  - 2.4.3 Evaluation . . . 25
- 2.5 Summary . . . 25

I The Role of Genres in Phrase-Based Machine Translation **27**

3 Genre and Topic Differences in Phrase-Based Machine Translation **29**
- 3.1 Introduction and research questions . . . 29
- 3.2 Domain adaptation for PBMT . . . 30
  - 3.2.1 Domain adaptation at the data level . . . 31
  - 3.2.2 Domain adaptation at the model level . . . 32
- 3.3 Genre and topic differences in PBMT . . . 33
  - 3.3.1 Definitions . . . 34
  - 3.3.2 Genre and topic adaptation . . . 35
  - 3.3.3 The Gen&Topic benchmark set . . . 35
- 3.4 The impact of genre and topic differences on PBMT . . . 36
  - 3.4.1 Translation quality . . . 36
  - 3.4.2 Model coverage analysis . . . 38
  - 3.4.3 Manual OOV analysis . . . 40
- 3.5 Conclusions . . . 42

4 Genre Adaptation Using Automatic Classifiers **43**
4.1 Introduction and research questions 43
4.2 Related work 44
4.2.1 Web as corpus 44
4.2.2 Text genre classification 45
4.2.3 Adaptation using classifiers 46
4.3 Bilingual resource acquisition 46
4.3.1 Language pairs and genres 47
4.3.2 Systematic data harvesting 48
4.3.3 Training and evaluation sets 49
4.4 Genre-specific baseline systems 50
4.4.1 Experimental setup 51
4.4.2 Results 53
4.5 Automatic genre classification 55
4.5.1 Building classifiers 56
4.5.2 Results 57
4.6 Genre adaptation using automatic classifiers 63
4.7 Conclusions 65

5 Genre Adaptation Using Text Features **67**
5.1 Introduction and research questions 67
5.2 Related work 68
5.2.1 Instance weighting for PBMT adaptation 68
5.2.2 Latent Dirichlet allocation 69
5.3 Translation model genre adaptation 70
5.3.1 VSM adaptation framework 70
5.3.2 Genre adaptation with subcorpus labels 71
5.3.3 Genre adaptation with genre features 72
5.3.4 Genre adaptation with LDA 73
5.4 Experimental setup 74
5.5 Results 75
5.5.1 VSM using intrinsic text features 75
5.5.2 VSM using manual subcorpus labels 77
5.6 Translation consistency analysis 78
5.7 Conclusions 79

II Translating and Analyzing Informal Language **81**

6 Translating and Analyzing User-Generated Text **83**
6.1 Introduction and research questions 83
6.2 Related Work 84
6.2.1 Analyzing PBMT errors 85
6.2.2 PBMT for informal language 85
6.3 Experimental setup 86
6.3.1 Evaluation sets 86

6.3.2 PBMT systems . . . 86
6.4 Error analysis and results . . . 87
6.4.1 Overall translation quality . . . 89
6.4.2 Translation phrase length analysis . . . 89
6.4.3 Model coverage analysis . . . 89
6.4.4 WADE: Word Alignment Driven Evaluation . . . 92
6.5 Conclusions . . . 96

7 Translating and Analyzing Fictional Dialogues 97
7.1 Introduction and research questions . . . 97
7.2 Related work . . . 98
7.2.1 Aligning and annotating fictional dialogue corpora . . . 99
7.2.2 Dialogue and discourse in PBMT . . . 99
7.3 Corpus construction and annotation . . . 100
7.3.1 Annotation process . . . 100
7.3.2 Post-processing and annotation quality . . . 101
7.4 Measuring dialogue effects on PBMT . . . 103
7.4.1 Basic experimental setup . . . 103
7.4.2 Approximate randomization testing . . . 105
7.5 Results . . . 106
7.5.1 The effect of dialogue acts on PBMT quality . . . 106
7.5.2 The effect of speakers on PBMT quality . . . 106
7.5.3 The effect of gender on PBMT quality . . . 108
7.5.4 The effect of register on PBMT quality . . . 109
7.6 Adaptation towards dialogue variables . . . 109
7.7 Conclusions . . . 111

III Domain Adaptation for Neural Machine Translation 113

8 Dynamic Data Selection for Neural Machine Translation 115
8.1 Introduction and research questions . . . 115
8.2 Related work . . . 116
8.2.1 Data selection for PBMT . . . 116
8.2.2 Domain adaptation for NMT . . . 117
8.2.3 Training efficiency for NMT . . . 118
8.3 Static data selection . . . 118
8.4 Dynamic data selection . . . 119
8.4.1 Sampling sentence pairs . . . 120
8.4.2 Gradual fine-tuning . . . 121
8.5 Experimental settings . . . 121
8.5.1 Machine translation systems . . . 121
8.5.2 Training and evaluation data . . . 122
8.6 Results . . . 122
8.6.1 Static data selection for PBMT and NMT . . . 122
8.6.2 Dynamic data selection for NMT . . . 125

8.7 Further analysis 128
8.8 Conclusions 130

9 Conclusions **131**
9.1 Main findings 131
9.2 Future work 135
9.2.1 Dynamic and fine-grained adaptation for MT 135
9.2.2 What NMT adaptation can learn from PBMT adaptation . . . 136

Bibliography **139**

Summary **151**

Samenvatting **153**

1 Introduction

Machine translation (MT) refers to the use of software to translate texts written in a *source* language (e.g., German) to a *target* language (e.g., English). Modern MT systems are typically statistical and data-driven, meaning that they are built using large quantities of data rather than hand-crafted rules. Before a data-driven statistical MT (SMT) system can be used for translation, it has to be trained on large amounts of example translations between the source and target language, so-called *parallel corpora*. Such bilingual corpora are publicly available for a select number of text types and language pairs, such as parliamentary proceedings in many European languages (Koehn, 2005). Consequently, translation output for such texts is often of good quality.

The picture looks less bright, however, when training data is scarce for the translation task at hand, for example if one wishes to translate TED talks or medical texts. In such cases, the majority of the available training data differs from the translation task in both writing style and vocabulary. This mismatch between the training data and the nature of the text to be translated can cause large drops in translation quality. To address this problem, a large body of previous work has targeted *domain adaptation*, in which an MT system is adapted to the *domain* of the translation task or test set, for instance by favoring the most relevant example translations (Axelrod et al., 2011; Costa-jussà and Banchs, 2011; Matsoukas et al., 2009; Foster et al., 2010; Chen et al., 2013, among others) or by combining models from different (*in-domain* and *out-of-domain*) MT systems (Banerjee et al., 2011; Bisazza et al., 2011; Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Sennrich, 2012b, among others).

Unfortunately, the notion of a domain is not uniformly defined among MT researchers. Typically, a domain is named after its origin or provenance, such as a particular website or a corpus number. This definition neglects the fact that *within* a single domain documents or sentences may vary at many levels. For example, different documents from the same website often discuss different *topics*, or belong to different *genres*, such as news articles or editorial pieces and their corresponding user comments. Large variations can also exist within genres, in particular between texts written by different authors lacking standardization, such as most user-generated (UG) data. Even more extreme examples are conversational texts, in which two consecutive utterances likely have different *speakers*, who each use a different style of language (reflected by *register*) and have different intents (reflected by *dialogue acts*).

To shed light on the concept of a domain and its impact on MT, the core question

in this thesis is: *"What's in a domain?"* To answer this question, we distinguish various aspects that together make up a domain, i.e., topic, genre, register, dialogue acts, speakers, and speaker gender, and we study to what extent MT output differs among these aspects, and how we can improve translation quality for each of the aspects. Adaptation to most of these fine-grained levels of a domain has gained little to no attention, and is thus a novel contribution of this thesis.

Within the large body of previous work on domain adaptation for MT, most research is driven by relatively formal genres such as news or parliamentary proceedings, which are characterized by a very standardized language use. These are, however, not representative for the abundance of informal or colloquial data emerging on the Internet, for which state-of-the-art MT systems perform markedly worse, a pattern observed in past MT evaluation campaigns. In this thesis, we pay special attention to analyzing and translating informal language. We distinguish in total six types of informal data: SMS messages, chat messages, telephone conversations, weblogs, user comments, and fictional dialogues. For each of these genres, we analyze which challenges they pose to MT and how translation quality can be improved.

In our experiments, we focus on domain or genre adaptation for MT without a dependency on corpus information. Most previous work on MT adaptation makes the strong assumption that the translation task has known domain or genre labels that are exploited to adapt the system. While this is a fair assumption in a controlled research scenario, it also limits the applicability of the proposed methods. In an online setting, for example, the genre or domain of the translation task is typically not known in advance. Throughout this thesis we present a number of approaches that move away from this assumption and apply genre or domain adaptation without relying on corpus labels.

Most of the analyses and experiments in this thesis are performed using *phrase-based* MT, to which we refer as PBMT. In Chapter 8, we also work with the recently developed paradigm of *neural* MT, to which we refer as NMT. We discuss both paradigms in detail in Chapter 2. While we cannot draw definitive conclusions for NMT based on findings in PBMT, and vice versa, we will end in Chapter 9 with a brief reflection on the lessons learned throughout this thesis and their expected implications on adaptation for NMT.

1.1 Research outline and questions

This thesis covers three research themes, all three related to domain adaptation in MT. The first theme concerns disambiguating the concepts *domain*, *genre*, and *topic* and studying their respective impacts on PBMT. Since genre differences appear to be particularly challenging to PBMT, we mainly focus on genre adaptation, for example by enhancing the bilingual training data with genre-specific resources, or by using genre-revealing textual features to improve translation.

The second theme deals with a particularly interesting genre: user-generated (UG) data. In an era of increasing use of the Internet, UG data is ubiquitous. However, it is also highly variable and not standardized or editorially controlled, making it difficult to translate automatically. Despite these challenges, most of the MT research to date has been driven by more formal translation tasks or by domains with very uniform language use, which are noticeably easier to translate. In this thesis we translate and

analyze a variety of colloquial and conversational genres and formulate challenges and recommendations for improving PBMT of these informal genres.

The third theme addresses domain adaptation for the new paradigm of NMT. Since domain adaptation for NMT is still an under-explored research direction, we use in this part of the thesis domains that are defined by their provenance rather than the fine-grained levels addressed in the first two research themes. Concretely, we apply a state-of-the-art adaptation approach to both PBMT and NMT and discuss the observed performance differences. Finally, we introduce a novel approach to improve upon the existing method for NMT adaptation.

Part I. The role of genres in phrase-based machine translation

An important goal of this thesis is to clarify the concept *domain* and its impact on PBMT. As a first step, we distinguish *genre* and *topic*—two different aspects that are often neglected when defining a domain by its provenance or corpus label—, and we ask:

RQ1 *What impact do genre and topic differences have on PBMT quality?*

- **a.** *Can we clarify the ambiguous use of the concept* domain *with regard to adaptation in PBMT?*
- **b.** *Which of two intrinsic text properties, topic and genre, presents a larger challenge to PBMT?*
- **c.** *To what extent do topic and genre differ with respect to out-of-vocabulary words and phrases?*

We answer RQ1 in Chapter 3 by (i) formulating clearer definitions of the concepts topic and genre, (ii) introducing a new benchmark set with controlled topic-genre distributions, and (iii) measuring and analyzing translation performance and model coverage variations at both the topic and the genre level.

Next, since we find that genre differences pose a large challenge to PBMT, we zoom in on genres alone. We investigate the impact of genre differences on PBMT in different language and resource scenarios, and we ask:

RQ2 *Is the observed impact of genre differences on PBMT consistent among various language pairs and data settings?*

- **a.** *To what extent does the impact of genre differences on PBMT vary among language pairs?*
- **b.** *To what extent can differences in PBMT performance among genres be explained by the proportion of genre-specific resources?*

We answer RQ2 in Chapter 4 by (i) automatically harvesting parallel training data for four genres in four language pairs, and (ii) analyzing correlations between translation performance and the proportion of language- and genre-specific resources.

After disentangling the concepts of genre and topic and analyzing their respective impact on PBMT, we move to adaptation. Most existing domain adaptation approaches

rely on the availability of domain labels for both the training and the test data. While this is a realistic assumption in controlled research scenarios, such labels may not be available when users consult a commercial MT system, or when (training) data is harvested from the web. In addition, even if all used corpora have known origin, genre, and topic, MT quality may still benefit from automatic detection of relevant data rather than relying on manually assigned labels. To address these issues, we investigate how to improve adaptation when corpus information is not available for the test data, the training data, or both. We therefore ask:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

- **a.** *Can we successfully adapt PBMT systems using automatic genre classifiers?*
- **b.** *Can we successfully adapt PBMT systems using textual indicators of genre?*
- **c.** *Can we successfully adapt PBMT systems using LDA features?*

We answer RQ3a in Chapter 4 by incorporating automatic genre classifiers into an end-to-end PBMT pipeline. We answer RQ3b and RQ3c in Chapter 5 by modifying a relevance weighting approach for domain adaptation such that it uses textual features or latent Dirichlet allocation (LDA) features rather than manually assigned subcorpus labels.

Part II. Translating and analyzing informal language

A genre that is particularly interesting in an era of increasing use of the Internet is user-generated (UG) data. While this term can be used to cover a variety of sub-genres, all UG texts have in common that they have been written by a lay-person, as opposed to a journalist or professional author, and that they have not undergone any editorial control. Consequently, most UG data is difficult to translate correctly. To gain a better understanding of the challenges of PBMT for UG data, we conduct a series of analyses on various UG benchmarks and ask:

RQ4 *How is translation quality of PBMT influenced by different types of user-generated text?*

- **a.** *What are the most common error types for various UG genres?*
- **b.** *To what extent do PBMT errors differ for different types of UG text and between UG text and news?*
- **c.** *What are promising strategies to improve PBMT for UG genres?*

We answer RQ4 in Chapter 6 by (i) quantitatively comparing PBMT performance of five UG benchmarks and two news data sets, and (ii) qualitatively analyzing common error types for these benchmarks.

An interesting informal genre that has only rarely been used in PBMT research, is conversational text. Conversations, or dialogues, involve—by definition—multiple speakers, and are thus noticeably different from more formal texts that are typically

written by a single writer with a single goal. Different speakers likely have different intentions and language use, demanding system adaptation at fine-grained levels of dialogue-specific aspects, such as *speakers*, *speaker gender*, *dialogue acts* and *register*. We study the impact of these aspects on PBMT quality and their potential to serve as indicators for adaptation by asking the following question:

RQ5 *What impact do dialogue-specific aspects have on translation quality?*

- **a.** *To what extent does PBMT quality vary between different speakers, speaker genders, dialogue acts and register?*
- **b.** *Which dialogue aspects are most indicative for PBMT quality?*
- **c.** *Can automatic annotations of dialogue-specific aspects benefit adaptation for PBMT?*

We answer RQ5 in Chapter 7 by (i) automatically extracting utterances from movie dialogues and annotating these utterances with speaker, speaker gender, dialogue act and register information, (ii) measuring translation quality variations for each of these aspects, and (iii) adapting PBMT systems to different dialogue acts and register levels.

Part III. Domain adaptation for neural machine translation

Finally, as research in MT is rapidly moving from PBMT to NMT, we are interested in how domain adaptation methods and findings from PBMT generalize to NMT. Compared to PBMT, research in domain adaptation for NMT is still in its infancy. We therefore start with a well-established PBMT method for domain adaptation, and investigate its applicability to NMT. The chosen method (Axelrod et al., 2011) performs data selection based on the training data's similarity to the domain of interest, and does not depend on the presence of corpus labels in the training data. For PBMT, this type of data selection improves domain-specific translation while simultaneously speeding up training. It would therefore be desirable if (a variant of) this method also succeeds for NMT. In summary, we ask:

RQ6 *To what extent and how can we successfully apply data selection for domain adaptation in NMT?*

- **a.** *How does a state-of-the-art data selection approach perform when applied to PBMT and NMT?*
- **b.** *Can we improve upon state-of-the-art data selection for NMT by dynamically changing the selected data subsets during training?*

We answer RQ6 in Chapter 8 by (i) comparing the performance of a state-of-the-art domain adaptation method for PBMT and NMT, and (ii) proposing two novel techniques to improve upon the existing approach for domain adaptation for NMT.

1.2 Main contributions

This thesis contributes at various levels to research in MT. Here we summarize the main algorithmic and empirical contributions, as well as the constructed resources.

Algorithmic contributions

Measuring model coverage: In Chapters 3 and 6 we propose and apply a method to measure PBMT model coverage for a given test set. Rather than only counting out-of-vocabulary (OOV) words, our metric covers *source phrase recall*, *target phrase recall* and *source-target phrase pair recall* for phrases and phrase pairs of any length.

Harvesting parallel data: In Chapter 4 we describe our step-by-step approach to automatic harvesting of parallel corpora from the web. This method is designed such that it can easily be used for a large number of web sources.

Genre adaptation using text features: In Chapter 5 we propose a way to improve upon an existing adaptation approach by replacing its dependency on subcorpus information with textual indicators of genre. The new method can be applied to any training corpus, even if it is not known from where the data originates.

Automatic detection and annotation of movie dialogues: In Chapter 7 we propose a procedure to (i) detect utterances and utterance boundaries from movie subtitles, and (ii) automatically annotate these utterances with speaker and gender information, dialogue acts and register levels.

Dynamic data selection for NMT: In Chapter 8 we propose dynamic data selection for NMT, for which we introduce two variants: *sampling* and *gradual fine-tuning*. Both techniques vary the size or composition of the selected subset of a parallel training corpus between training epochs. While sampling yields varying results, gradual fine-tuning performs substantially better than static data selection.

Empirical contributions

Genre and topic differences in PBMT: In Chapter 3 we formulate clear definitions of the notions *genre* and *topic*, which together make up a domain. The definitions are based on research literature in the fields of text analysis and classification. Next, we show that genre differences pose a larger challenge to PBMT than topic differences, and that this difference can be largely attributed to poor model coverage.

Resources versus language pairs: In Chapter 4 we show that PBMT quality differences between genres generalize to a large degree across language pairs, and that these differences can partially be explained by the proportion of genre-specific bilingual resources in the training data.

Genre adaptation using automatic classifiers: In Chapter 4 we show that PBMT quality for a test set comprising a mixture of genres can be improved substantially when employing genre classifiers that route each test document to the most appropriate genre-specific PBMT system. We compare various types of classifiers and feature sets, and show that using a support vector machines (SVM) classifier in combination with bag-of-words (BOW) features performs best on this task.

Comparing user-generated (UG) data sets: In Chapter 6 we show that different types of UG data exhibit different PBMT errors, but also that there are some common patterns among various UG test sets. For example, we find that UG data is translated with shorter phrases than news, which can likely be explained by poor bilingual model coverage.

Analyzing dialogue aspects: In Chapter 7 we show that BLEU fluctuations between speakers, speaker genders, dialogue acts and register values are more prominent than randomly expected, and that these aspects can be used to successfully adapt PBMT systems.

Static data selection for PBMT and NMT: In Chapter 8 we show that a state-of-the-art data selection method that works consistently well for PBMT is much less effective for NMT, requiring much larger selections to achieve similar results.

Resources

Gen&Topic data set and OOV annotations: In Chapter 3 we introduce an Arabic-English development and test set, the Gen&Topic benchmark, which have controlled topic-genre distributions. In addition, we annotate observed OOVs in the Gen&Topic test set according to an OOV taxonomy distinguishing five main OOV classes.

Web-harvested parallel genre bitexts: In Chapter 4 we introduce new parallel training, evaluation and test data covering four genres (colloquial, editorial, news, and speech) for four language pairs (Arabic-English, Bulgarian-English, Chinese-English, and Persian-English).

Movie dialogue data set: In Chapter 7 we introduce a data set covering five language pairs with utterances extracted from movie subtitles. All utterances are annotated with speaker, speaker gender, dialogue act, and register labels.

1.3 Thesis overview

After this introductory chapter, the remainder of this thesis consists of a background chapter (Chapter 2), six research chapters (Chapters 3–8) covering three research themes, and a concluding chapter (Chapter 9). Below we present a high-level overview of the main content of each of these chapters. Each part covering a research theme is self-contained and can be read without knowledge of the other parts.

Chapter 2: Background provides an introduction to the two MT paradigms used in this thesis: phrase-based MT (PBMT) and neural MT (NMT). For both paradigms, we briefly discuss the core models, the training data needed, and the learning and optimization strategies we employ. Moreover, we describe the basic experimental settings for our in-house PBMT system Oister and NMT system Tardis.

Part I. The role of genres in phrase-based machine translation

In the first part of this thesis we shed light on the concepts *domain*, *genre*, and *topic*, and we introduce two methods for genre adaptation.

Chapter 3: Genre and Topic Differences in PBMT disentangles the concepts *genre* and *topic* and studies their respective impact on PBMT quality. We first introduce clear definitions of both concepts, then present the Gen&Topic benchmark, which has controlled topic-genre distributions, and finally use this data set to perform an extensive analysis of word and phrase pair coverage for different genres as well as topics.

Chapter 4: Genre Adaptation Using Automatic Classifiers introduces genre adaptation using automatic genre classifiers. To this end, we first describe our efforts to harvest from the web large amounts of parallel data in four genres and four language pairs. After further analyzing the impact of genre differences on PBMT quality, we train genre classifiers which accurately classify our test documents in five languages. Finally, we incorporate these classifiers into an end-to-end PBMT pipeline for successful genre adaptation.

Chapter 5: Genre Adaptation Using Text Features introduces genre adaptation using textual features and LDA. We improve upon an existing domain adaptation approach by replacing its dependency on subcorpus information with textual indicators of genre.

Part II. Translating and analyzing informal language

In the second part of this thesis we analyze the impact of informal language such as found in SMS messages, user comments or dialogues on PBMT quality.

Chapter 6: Translating and Analyzing User-Generated Data analyzes the PBMT errors for five types of user-generated (UG) genres, both quantitatively and qualitatively. Concretely, we compare five UG and two news data sets, to discover both differences and consistencies among various types of UG data. For this study, we introduce novel error analysis measures and adopt existing ones.

Chapter 7: Translating and Analyzing Fictional Dialogues analyzes the impact of conversational aspects on PBMT quality. First, we create a movie dialogue corpus by annotating parallel subtitles with *speaker*, *speaker gender*, *register*, and *dialogue act* information. Next, we study to what extent translation quality differs for each of these aspects. Finally, we investigate whether PBMT benefits from a simple adaptation approach at fine-grained dialogue-specific levels.

Part III. Domain adaptation for neural machine translation

In the third part of this thesis we address domain adaptation for NMT. Here we ignore more fine-grained aspects of domains since adaptation for NMT is still in its infancy.

Chapter 8: Dynamic Data Selection for NMT investigates the effectiveness of a state-of-the-art data selection method for domain adaptation for both PBMT and NMT, and introduces two variants of *dynamic data selection* to make data selection profitable for NMT.

Finally, we summarize findings from all three research themes in the concluding chapter:

Chapter 9: Conclusions concludes this thesis by recapitulating the research questions and their respective answers. We also reflect on interesting future research directions and on what NMT can learn from findings in PBMT and findings in this thesis.

1.4 Origins

The research presented in Chapters 3–8 of this thesis is based on a number of peer-reviewed publications. Below we indicate the origins of each chapter and we specify the roles of the authors.

Chapter 3 is based on van der Wees, Bisazza, Weerkamp and Monz (2015c), *What's in a Domain? Analyzing Genre and Topic Differences in Statistical Machine Translation*, published in ACL 2015 (Short Papers). Van der Wees created the Gen&Topic data set and designed and carried out the MT experiments. Bisazza performed the manual analysis of unseen Arabic words. All authors contributed to the text.

Chapter 4 is based on technical project reports and comprises work on harvesting parallel corpora from the web, learning to classify texts by genre, and adapting PBMT systems using automatic genre classifiers. Van der Wees designed the methods, performed the experiments, and wrote the text in this chapter.

Chapter 5 is based on van der Wees, Bisazza, and Monz (2015b), *Translation Model Adaptation Using Genre-Revealing Text Features*, published in the Second Workshop on Discourse in Machine Translation (DiscoMT 2015). Van der Wees designed the methods, performed the MT experiments and wrote most of the text.

Chapter 6 is based on van der Wees, Bisazza, and Monz (2015a), *Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text*, published in the First Workshop on Noisy User-generated Text (W-NUT 2015). Van der Wees designed the methods, performed the PBMT experiments and wrote most of the text. Bisazza manually analyzed the Arabic examples. This paper was awarded the best paper award at the workshop.

Chapter 7 is based on van der Wees, Bisazza, and Monz (2016a), *Measuring the Effect of Conversational Aspects on Machine Translation Quality*, published in COLING 2016. Van der Wees created the annotated movie dialogues corpus, designed and performed the MT methods and experiments and wrote most of the text.

Chapter 8 is based on van der Wees, Bisazza, and Monz (2017), *Dynamic Data Selection for Neural Machine Translation*, published in EMNLP 2017. Van der Wees developed the methods, performed the PBMT and NMT experiments and wrote most of the text.

Finally, work in this thesis benefits from results published in van der Wees, Bisazza, and Monz (2016b), *A Simple but Effective Approach to Improve Arabizi-to-English Statistical Machine Translation*, published in the Second Workshop on Noisy User-generated Text (W-NUT 2016). For this work, an Arabizi (i.e., Romanized Dialectal Arabic) transliteration software component and an Arabizi-English parallel data set were released.

2 Background

In this chapter we discuss the two machine translation (MT) paradigms that are used in this thesis: phrase-based machine translation (PBMT, Section 2.2) and neural machine translation (NMT, Section 2.3). Both are *statistical* approaches to automatic translation, in which the models are generated from large amounts of data using machine learning techniques. Since the training data is an essential ingredient for building statistical models, we first devote Section 2.1 to describing this data. Finally, in Section 2.4 we detail the main experimental setup used in the research chapters of this thesis.

Naming conventions. Statistical machine translation is generally referred to as SMT and includes a number of approaches such as PBMT, hierarchical MT (Chiang, 2007), and syntax-based MT (Zollmann and Venugopal, 2006). Throughout this thesis we focus on the SMT paradigm of PBMT, which is one of the most widely used approaches. While NMT is also a statistical approach to MT, it is typically not referred to as SMT. To maximize clarity, we use in this thesis the terms PBMT and NMT when explicitly addressing phrase-based or neural MT, respectively. We only refer to (S)MT when making general statements about (statistical) machine translation.

2.1 Parallel and monolingual corpora

Both PBMT and NMT are *data-driven* and *statistical*: their model components are trained on large collections of data in a statistical manner. Since the goal of MT is to automatically translate text in a *source* language (e.g., German) into a *target* language (e.g., English), the most important input data comprises example translations between the two languages of the *language pair* of interest. A collection of such example translations is called a *parallel corpus* or *bitext*. Typically, example translations in a bitext are paired at the sentence level, yielding parallel *sentence pairs*. How many sentence pairs are needed for good translation results depends to a large extent on the language pair and the nature of the data, but a complete bilingual training corpus often contains hundreds of thousands up to several millions of sentence pairs (Bojar et al., 2013, 2014, 2015, 2016).

Besides quantity, SMT performance is very dependent on the quality of the training data and its suitability for the translation task of interest. Regarding quality, optimal

parallel corpora consist of professional translations, however these are expensive and often not easy to obtain. Professionally translated corpora are limited to parliamentary proceedings such as Europarl (Koehn, 2005) or the Canadian Hansard corpus,[1] and a number of smaller corpora available through the Linguistic Data Consortium (LDC).[2] If one wishes to translate texts of a different nature, it is not uncommon to use non-professional or even automatic translations that are of lower quality but more relevant for the translation task (Bertoldi and Federico, 2009; Sennrich et al., 2016b).

While parallel corpora are a necessary input to train any type of SMT system, PBMT systems also employ *monolingual* corpora. Such corpora contain texts in the target language and are used to increase the fluency of the output translations. Since monolingual corpora do not require manual translations, they are much more abundant and easier to obtain than parallel corpora. A typical monolingual corpus used for PBMT easily contains over a hundred million sentences (Bojar et al., 2013, 2014).

2.2 Phrase-based machine translation

Phrase-based machine translation (PBMT) has been the state-of-the-art approach to automatic translation since the early 2000s (Koehn, 2009). As the name suggests, phrase-based MT uses *phrases* as its main translation units. Phrases are contiguous sequences of words, usually up to a pre-defined maximum number of words, that do not need to have linguistic coherence but instead appear in the training bitext. This is for example illustrated in Figure 2.1b, where valid extracted phrase pairs include '*animals are in* ↔ *Tiere leben in*' and '*are in the* ↔ *leben in den*'.

The PBMT approach uses a log-linear combination of several models capturing different aspects of translation: The *translation model* takes care of the translation options, the *reordering model* deals with word order differences between the source and the target language, and the *language model* ensures fluency of output translations. Below we discuss each of these models (§2.2.2–2.2.4), followed by an explanation of their parametrized log-linear combination used for translation (§2.2.5). In the next section, we first tackle the tasks of word alignment and phrase extraction, which are necessary steps for the creation of translation and reordering models.

2.2.1 Word alignment and phrase extraction

Given a parallel training corpus, the first goal is to discover which words are translations of each other. This is done by automatically inferring hidden 'links' between words in the source and target language, called *word alignments*. Even if one would have access to a traditional dictionary, two challenges complicate the process of inferring word alignments: First, words are inherently ambiguous and can often translate into various words in another language, depending on the context. Second, sentences in different languages often differ in their word order, making word alignment a non-monotonic process and dramatically increasing the number of possible alignment links.

[1] LDC corpus number LDC95T20.

[2] `https://www.ldc.upenn.edu/`

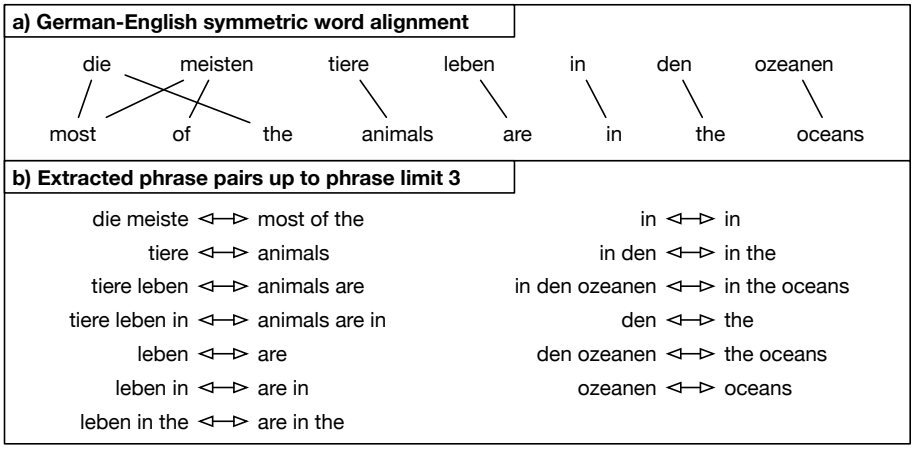


Figure 2.1: a) German-English sentence pair with its word alignment. b) the sentence pair's extracted phrase pairs using a phrase limit of 3. Note that for example the translation '*meiste* ↔ *most*' cannot be extracted from this particular sentence pair since both words have alignment links with other words and would thus violate the consistency constraints.

Word alignment is typically done using IBM models 1–5 (Brown et al., 1993) and HMM-based alignment (Vogel et al., 1996), which create one-to-many alignments in two language directions (source-target and target-source). Search for the most plausible alignment links for each direction is done using the Expectation Maximization (EM) algorithm (Dempster et al., 1977). Merging both alignments results in a single *symmetric* many-to-many alignments, in which words can also stay unaligned. A common method of merging alignments and adding neighboring unaligned words is the *grow-diag-final-and* heuristic (Koehn et al., 2005). Word alignment quality depends on reliable statistics and is usually poor when using insufficient amounts of training data. An example of a word alignment for a German-English sentence pair is shown in Figure 2.1a.

Next, from the symmetric word alignments we can extract source-target *phrase pairs*. In the phrase extraction process, all phrase pairs that are *consistent* with the word alignment are collected. To be consistent, a candidate phrase pair (i) only includes consecutive sequences of words maintaining their original order, (ii) contains at least one word alignment link, and (iii) has no alignment links with words *outside of* the phrase pair. Figure 2.1b illustrates all the phrase pairs that can be extracted from the word alignment in Figure 2.1a. Once all phrase pairs for a given bitext have been collected, we can compute each phrase pair's translation and reordering probabilities, which we discuss in the next two sections.

2.2.2 Translation models

Translation models, also called *phrase tables*, can be considered as large dictionaries containing phrase translations with their corresponding translation probabilities. These probabilities are estimated by counting and normalizing occurrences of phrase and word pairs in the training bitext. Formally, the conditional probability of translating source

phrase $\bar{f}$ into target phrase $\bar{e}$ is estimated as the relative frequency of joint counts:[3]

$$p\left(\bar{f} \mid \bar{e}\right) = \frac{c\left(\bar{e}, \bar{f}\right)}{c\left(\bar{e}\right)}. \tag{2.1}$$

Here, $c\left(\bar{e}, \bar{f}\right)$ is the joint count of $\bar{e}$ and $\bar{f}$ in the training bitext, and $c\left(\bar{e}\right)$ is the total count of $\bar{e}$. Phrase translation probabilities for the inverse translation direction $p\left(\bar{e} \mid \bar{f}\right)$ can be computed similarly.

Unfortunately, phrase translation probabilities can be unreliable, especially for infrequent phrases. To overcome this problem, it is common to decompose phrase pairs and compute *lexical* translation probabilities for the aligned words within a phrase pair (Koehn et al., 2003). Lexical translation probabilities are estimated as:

$$p_{\text{lex}}\left(\bar{f} \mid \bar{e}\right) = \prod_{i=1}^{|\bar{e}|} \frac{1}{|\{j \mid (i,j) \in a\}|} \sum_{\forall (i,j) \in a} p_w\left(e_i \mid f_j\right), \tag{2.2}$$

where a refers to the phrasal word alignment, and $p_w\left(e_i \mid f_j\right)$ are word translation probabilities for target word e_i given its aligned source word f_j, computed following (2.1). Inverse lexical translation probabilities are computed similarly.

While phrase and lexical translation probabilities are the core components of translation models, it is common to also include a *word penalty* and a *phrase penalty*. These are factors that encourage a bias towards using either fewer or more words or phrases, respectively, during translation. The word penalty factor ω favors the use of fewer words, i.e., shorter translations, if $\omega < 1$ and more words, i.e., longer translations, if $\omega > 1$. The phrase penalty factor ρ favors the use of fewer phrases, i.e., longer average phrase length, if $\rho < 1$, and more phrases, i.e., shorter average phrase length, if $\rho > 1$. Word and phrase penalties do not depend on the actual phrase pairs, but are general system features.

Besides word and phrase translation probabilities, other features can be added to translation models. In Chapter 5, for example, we add features capturing phrase-pair-specific information used for genre adaptation.

2.2.3 Reordering models

Like translation models, reordering models contain source phrases with their potential translations in the target language. However, rather than providing information on which translation to generate, reordering models contain information on the *order* in which phrase translations should be generated. Each phrase pair in the reordering model contains probabilities for its orientation with respect to the previous and the next phrase, considering three orientation types: monotone, swap, and discontinuous. It is common to use *lexicalized* reordering models (Tillmann, 2004), in which the orientation probabilities are conditioned on the actual phrases and their counts in the training bitext.

Besides orientation probabilities, reordering behavior of PBMT is also affected by the *distortion cost* and the *distortion limit* (Berger et al., 1996). The distortion cost adds

[3] It is historically common to denote the two languages of a language pair with F (for *foreign* or *French*) and E (for *English*), indicating the source and target, respectively.

a penalty for the length of each non-monotonic reordering, and the distortion limit indicates the maximum length of a non-monotonic 'jump' to allow during translation. This limit helps reducing translation complexity but also forbids long-distance reorderings, which is undesirable for some language pairs. A comprehensive survey on reordering models is provided by Bisazza and Federico (2016).

2.2.4 N-gram language models

The language model (LM) is the only PBMT model trained on monolingual data. By observing word and phrase counts in a large monolingual corpus in the target language, the LM estimates probabilities reflecting the likelihood of a sequence of words to be uttered by a fluent English speaker. While their main function in PBMT is to ensure fluency of output translations, they also help making translation and word order decisions. A good LM favors (i) correct word order over incorrect word order:

$$p_{\text{LM}}(\texttt{the cat meows}) > p_{\text{LM}}(\texttt{the meows cat}),$$

and (ii) common word combinations over rare word combinations:

$$p_{\text{LM}}(\texttt{the cat meows}) > p_{\text{LM}}(\texttt{the cat barks}).$$

The language models used in PBMT are usually *n-gram* LMs, modeling contiguous sequences of words up to length n. In n-gram LMs, probabilities are estimated from n-gram counts in the training corpus. For example, for a *trigram* LM, we estimate the likelihood of observing word w_3 directly after observing words w_1 and w_2 as follows:

$$p(w_3 \mid w_1, w_2) = \frac{c(w_1, w_2, w_3)}{\sum_{w_i} c(w_1, w_2, w_i)}, \tag{2.3}$$

where $c(w_j, w_k, w_l)$ is the raw count of string "$w_j\ w_k\ w_l$" in the training corpus. To compute the probability of an entire sentence s, the LM multiplies the observed probabilities[4] of each word given its history of $n-1$ words:

$$p(s) = \prod_{i=1}^{|s|+1} p(w_i \mid w_{i-n+1}, \ldots, w_{i-1}). \tag{2.4}$$

LM probabilities also include the tokens ⟨s⟩ and ⟨/s⟩, indicating the beginning and end of a sentence, respectively. Using these, an LM can estimate the likelihood of a given word to start or end a sentence.

The above equations ignore an important problem: If a given n-gram in the output sentence under translation has never been observed in the training corpus, its LM probability equals zero, making the probability of the entire sentence zero as well. This problem is typically solved by backing off to lower order n-grams and by applying *smoothing* techniques that add a small probability mass to unseen events. A comprehensive empirical comparison of LM smoothing techniques is presented by Chen and Goodman (1999).

[4]To avoid having to compute extremely small numbers, it is in practice more common to add up negative log-probabilities rather than to multiply probabilities.

2.2.5 Decoding and the log-linear model

Decoding is the process of finding the best automatic translation of an incoming source sentence. During decoding, the PBMT system (or *decoder*) uses a parametrized log-linear combination (Och and Ney, 2002) of the different models discussed in the previous sections. Formally, the aim of the decoder is to find the most probable target sentence $\hat{E}$ given a source sentence F by scoring possible translation hypotheses:

$$\hat{E} = \arg\max_{E,a} \exp \sum_{i=1}^{n} \lambda_i h_i \left(F, E, a\right). \tag{2.5}$$

Here, a is a latent variable representing phrasal alignments between F and the translation hypothesis under consideration E, $h_i \left(F, E, a\right)$ are n feature functions, and λ_i their corresponding feature weights. Typically, feature functions are the log-probabilities of the above discussed translation, reordering and language models as well as the features word penalty, phrase penalty, and distortion cost.

Parameter optimization

Finding the optimal values for the feature weights λ_i is done during parameter optimization, commonly referred to as *tuning*. During tuning, translations are generated for a held-out or *development* data set, and compared to the set of reference translations of this development set, thus directly optimizing translation quality. Two of the most common parameter optimization techniques are minimum error rate training (MERT, Och (2003)) and pairwise ranking optimization (PRO, Hopkins and May (2011)). The final set of feature weight values is highly dependent on the nature of the development set. It is therefore common practice to use a development set that is representative of the translation task in terms of vocabulary and style. Using a development set that differs too much from the desired translation task likely results in decreased translation performance.

Decoding process

The actual decoding process consists of (i) decomposing the source sentence into phrases that occur in the phrase table, (ii) finding a suitable translation for each source phrase, and (iii) placing these phrase translations in the optimal order. For each sentence there are many different phrase segmentations, many translation options, and many reordering possibilities. Searching through the entire space of possible translation hypotheses to find the single best translation is computationally infeasible, and PBMT decoding has been shown to be NP-complete (Knight, 1999). PBMT decoding is therefore performed by approximate search methods employing a number of different techniques, which we discuss below and illustrate with a translation example in Figure 2.2.

Translation hypotheses are generated from left to right. Starting from an empty hypothesis, phrase translations can be added for different phrase segmentations of the source sentence. Note that the source sentence does not have to be processed from left to right: phrases can be reordered provided that the distortion limit constraint is never violated. During the decoding process, a coverage vector keeps track of which

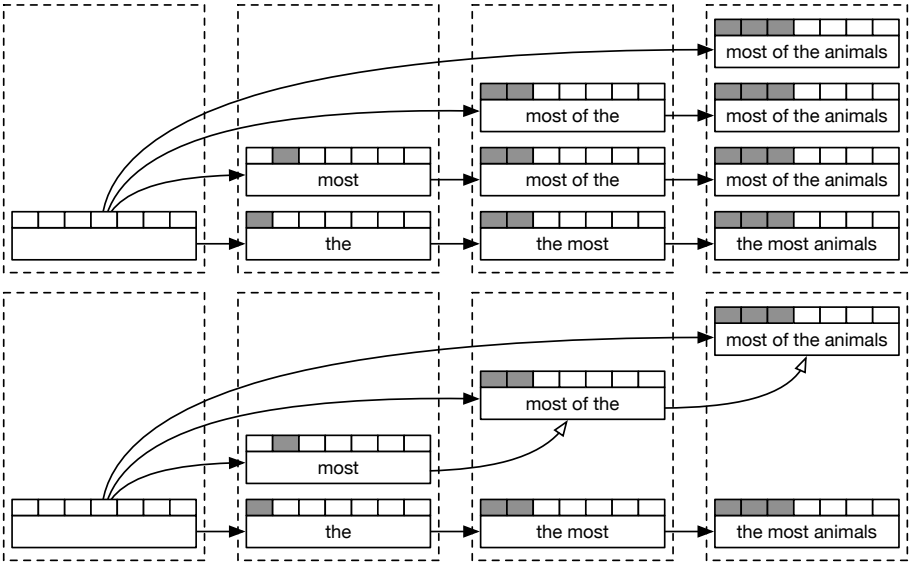


Figure 2.2: Example of PBMT stack decoding. Top: translation hypothesis expansion of the German input sentence 'Die meisten Tiere leben in den Ozeanen'. The coverage vector indicates which source words have been translated. Bottom: recombination of identical partial hypotheses (light arrows). Typically, hypotheses with more (and shorter) phrase translations are discarded.

source words have been translated so far, and ensures that eventually each source word is translated exactly once (Nießen et al., 1998). Each partial hypothesis is stored in a *stack* together with other hypotheses with the same number of translated source words (Tillmann et al., 1997).

To increase search efficiency, similar translation hypotheses with a different construction history (e.g., different phrase segmentations) can sometimes be merged, a process called *recombination* (Zens et al., 2002). The idea behind recombination is that a hypothesis with a lower probability than another hypothesis that is identical from a search perspective can never become the best-scoring hypothesis anymore, and can thus be discarded. Two partial hypotheses can be recombined if they meet a number of criteria: the hypotheses must have (i) identical coverage vectors, (ii) the same last $n-1$ words (such that the language model cannot distinguish the two hypotheses), and (iii) the same position of the most recently translated phrase (such that the reordering models cannot distinguish the two hypotheses).

Unfortunately, recombination alone cannot sufficiently decrease decoding complexity. Another, more radical solution to decreasing the search space is to *prune* out hypotheses that are not considered good enough (Wang and Waibel, 1997). This is done, for example, by only allowing a maximum number of hypotheses per stack (*histogram* pruning), or by discarding hypotheses whose probability is more than a threshold α times worse than the best hypothesis in the same stack (*threshold* pruning). Pruning-based decoding is also referred to as *beam search* since it resembles a beam of light shining on the best hypothesis and only illuminating other hypotheses that are in

close vicinity of the best one. In beam search, the *beam width* is used ambiguously to indicate either the number of hypothesis to keep or the pruning threshold α to use.

Pruning does not come without risks. Based on partial translations, one could prune out hypotheses that may eventually have been extended to the optimal translation. This risk to make *search errors* is exacerbated by the fact that partial hypotheses in the same stack can have different coverage vectors and that some partial translations may have used easier phrase segmentations than others. The second stack in Figure 2.2 shows an example of two hypotheses in the same stack with different spans of translated source words. When deciding which hypotheses to prune, we do not only want to consider the probability score of the hypothesis, but also its expected *future cost*. Since computing the exact future cost is computationally intractable, we instead estimate how difficult the remaining part of the sentence is to translate. Typically this is done using an optimistic combination of the translation and language model scores for the phrases still to be translated. Detailed explanations and derivations of future cost estimation and all other discussed models are provided by Koehn (2009).

2.3 Neural machine translation

End-to-end neural machine translation (NMT) is a novel machine translation paradigm in which translation is performed by a neural network (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014b). Also known as the *encoder-decoder* framework, this complex neural model performs translation as a bilingual language modeling task predicting output words one by one. To this end, it first *encodes* a source sentence as a vector and then *decodes* this vector to an output sentence in the target language. Unlike in PBMT, no explicit modeling of reordering, phrase segmentation, or word alignment is needed in NMT, making it a conceptually simpler MT paradigm.

In this section we first discuss at a high level the basic components of an NMT system: word embeddings (§2.3.1), recurrent neural language models (§2.3.2), the encoder-decoder architecture (§2.3.3), and the attention mechanism (§2.3.4). Next, we explain how to train and optimize an NMT system (§2.3.5) and how to use it for translation (§2.3.6).

2.3.1 Word embeddings and vocabulary size

Unlike PBMT, which uses a large monolingual corpus for language modeling, NMT requires as its training data only a parallel corpus. Starting from this bitext, the first step is to convert both the source and the target tokens to numerical representations. This is done using *word embeddings*, which represent words as vectors of real numbers, based on the observed contexts of a word in the training corpus (Collobert et al., 2011). In theory, elements in the embedding vectors can reflect certain aspects of words, for example a word being feminine or masculine. While in practice it is rare for vector elements to have this kind of intuitive meaning, words with similar characteristics typically have similar embedding vectors. We can even perform arithmetic operations with embedding vectors (Mikolov et al., 2013b; Levy and Goldberg, 2014).

While word embeddings can be pre-trained on monolingual corpora (Mikolov et al., 2013a; Pennington et al., 2014) before being used for translation, most end-to-end NMT systems learn the embedding vectors as part of the training process (see Section 2.3.5).

For a given training corpus, the number of words in the vocabulary can be extremely large (easily over a million, depending on the corpus size and the morphological complexity of the language), with many words occurring only once or a few times. Since these words are likely to have unreliable statistics, and to reserve some representation for words that do not appear at all in the training data, it is common to restrict the vocabulary size to a maximum of V words (commonly between 30K and 80K) and replace rare words with a special token representing unknown words: $\langle \text{UNK} \rangle$ (Sutskever et al., 2014).

2.3.2 Recurrent neural language models

The main building blocks of an NMT system are *recurrent neural networks* (RNNs, Mikolov (2010)). Like other types of neural networks, RNNs take a numerical input, apply a non-linear operation in a *hidden* layer, and predict an output value. However, unlike simpler (feed-forward) neural networks, RNNs are characterized by having connections between hidden states of sequentially connected network units, making them very suitable for modeling sequential events such as strings of words. For example, RNNs can be used as a language model that predicts the most likely next word given a sequence of preceding words. In this scenario, the main input of an RNN architecture are embedding vectors of words, and the predicted output is a probability distribution over a vocabulary V of words. More formally, at each time step $t \geq 1$, the input of the network's hidden states $\boldsymbol{h}_t$ is defined by the embedding vector $\boldsymbol{m}_t$ of the current word as well as by the output of the hidden state $\boldsymbol{h}_{t-1}$ of the RNN at time step $t-1$. The advantage of this hidden state connection is that information from previous parts of the sentence is automatically passed on to the next RNN unit, and prediction of each word can be conditioned on an arbitrarily long history.

RNNs come in different flavors. One of the most commonly used RNNs is the *long short-term memory* (LSTM, Hochreiter and Schmidhuber (1997); Gers et al. (2000)). Compared to the above described general RNN setup, LSTMs explicitly model a *memory*. This memory consists of a number of *gates* that manage the information flow through the network. Typically, an *input gate* controls which information to allow into the memory, an *output gate* controls which information that is stored in the memory to use, and a *forget gate* controls which information to keep in memory and which to forget. Thanks to their memory architecture, LSTMs are superior to standard RNNs (and even more so to feed-forward neural LMs or n-gram LMs) in capturing semantic and syntactic long-distance dependencies (Neubig, 2017).

2.3.3 Encoder-decoder models

The encoder-decoder framework used for end-to-end NMT is similar to the above described LSTM (Sutskever et al., 2014). However, instead of computing the probability $p(E)$ of sentence E, NMT models the probability $p(E \mid F)$ of target sentence E given source sentence F.

A schematic overview of an encoder-decoder model is shown in Figure 2.3. As suggested by the name, this model consists of two components: an encoder and a decoder. The encoder processes an incoming source sentence F word by word. Each word is converted to an embedding vector before being fed as input to the network. In addition, hidden states capturing information of previous words are passed on to the current time step, and when the entire source sentence has been processed, the last hidden state contains information about the entire sentence. It is common practice to encode the source sentence in reverse order such that the first word of the source sentence is remembered most clearly by the network, which has been shown to improve translation quality (Sutskever et al., 2014; Luong et al., 2015a). The intuitive explanation is that for many language pairs the beginning of the source sentence is likely the most relevant part for translating the beginning of the target sentence. However, if the word ordering between the source and target language is very different, it is more beneficial to use a *bidirectional* encoder (Bahdanau et al., 2014).

The second component of the NMT architecture, the decoder, takes as its input the last hidden state of the encoder and an embedding vector for the special start-of-sentence token $\langle \mathrm{s} \rangle$. The network then computes conditional probabilities[5] over words in a fixed target language vocabulary $e \in V_{\mathcal{E}}$:

$$p\left(e_t \mid e_0^{t-1}; F\right). \tag{2.6}$$

For each time step t, the probabilities over $V_{\mathcal{E}}$ depend on the history of predicted target words and on the complete source sentence. Decoding stops when a special end-of-sentence token $\langle /\mathrm{s} \rangle$ is generated.

State-of-the-art encoder-decoder models contain multiple (typically four) layers of stacked LSTMs, yielding more powerful models than single-layer LSTMs. The observed improved performance using multiple layers may be explained by the fact that different layers learn to encode different types of linguistic information (Shi et al., 2016). Figure 2.3 illustrates a two-layer stacked LSTM.

2.3.4 Attention mechanisms

The NMT architecture discussed so far has an important shortcoming: the hidden state vectors capturing the source sentence have a pre-determined fixed size, regardless of the length of the input sentence. As a consequence, the system cannot accurately store all of the information for long sentences, leading to deteriorated translation quality. In addition, the encoder-decoder model does not exploit the fact that some source words are more relevant than others for the prediction of a given target word.

To overcome this limitation, current state-of-the-art NMT systems employ an *attention mechanism* (Bahdanau et al., 2014; Luong et al., 2015a), which tells the system how much to "focus" on each source word when predicting a particular target word. Concretely, the attention mechanism includes an *attention vector* $\boldsymbol{\alpha}_t$ that contains scores representing the similarity of each hidden state of the source sentence $\boldsymbol{h}_j^{(f)}$ with the hidden state of the current target word $\boldsymbol{h}_t^{(e)}$. A common way to compute the attention

[5] Probabilities are computed using a softmax function over the network output, see Section 2.3.5.

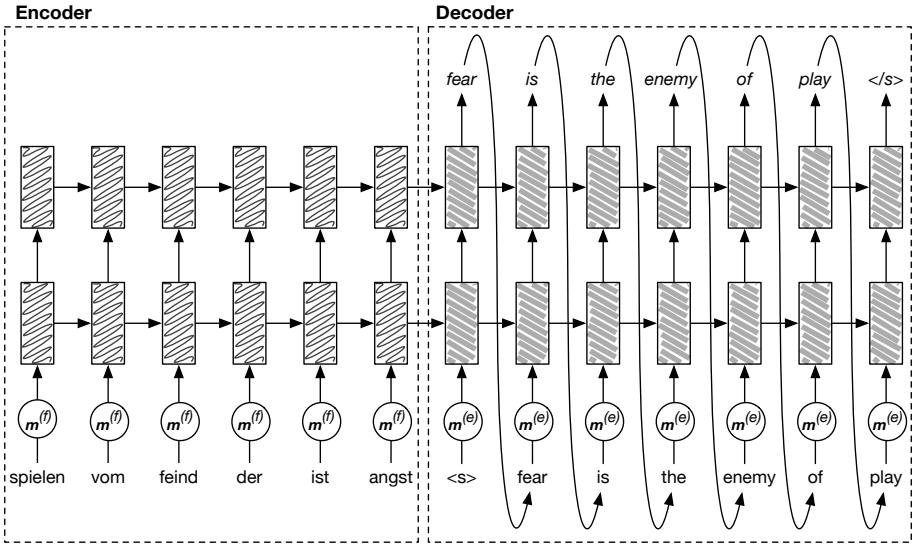


Figure 2.3: Schematic representation of an encoder-decoder model with two stacked layers of LSTMs. The input sentence 'Angst ist der Feind vom Spielen' is processed in reverse order. Circles with $\boldsymbol{m}^{(f)}$ and $\boldsymbol{m}^{(e)}$ represent embedding lookup tables for source and target words, respectively.

vector is by first computing the dot product of the hidden states of target word e_t and source word f_j:

$$a_{t,j} = {\boldsymbol{h}_t^{(e)}}^\top \boldsymbol{h}_j^{(f)}. \tag{2.7}$$

Combining these scores for all source words $j \in \{1, \ldots, m\}$ yields a score vector $\boldsymbol{a}_t$ for time step t. We obtain the attention vector by taking the softmax over $\boldsymbol{a}_t$:

$$\boldsymbol{\alpha}_t = \text{softmax}\left(\boldsymbol{a}_t\right). \tag{2.8}$$

Alternative functions to compute $\boldsymbol{a}_t$ are discussed by Luong et al. (2015a).

Next, the attention vector is used to derive a context vector $\boldsymbol{c}_t$ from the hidden source sentence representations $\boldsymbol{h}_j$:

$$\boldsymbol{c}_t = \sum_{j=1}^{n} \alpha_{t,j} \boldsymbol{h}_j^{(f)}. \tag{2.9}$$

This context vector is taken into account when making target word predictions during translation, which we show in (2.11) in the next section.

Attention has proven a very powerful mechanism, and several variants exist. Some approaches also use *input feeding* (Luong et al., 2015a), in which the attentional vector $\boldsymbol{\alpha}_t$ is fed to the input of the next hidden state $\boldsymbol{h}_{t+1}$. In fact, it has recently been argued that 'attention is all you need' to achieve competitive translation performance (Vaswani et al., 2017). While attention resembles traditional word alignment, the two often do not correspond (Koehn and Knowles, 2017; Ghader and Monz, 2017).

2.3.5 Training an NMT system

The performance of an NMT system depends to a large extent on the values of the model parameters. While in PBMT it is common to first train the various model components and then optimize their feature weights, NMT training consists of directly optimizing all parameters on the training bitext. When the basic model architecture (e.g., number of layers, embedding vector size, hidden state vector size) has been determined, the only task is to find the optimal set of parameter values. However, the number of parameters in NMT is extremely large, requiring advanced optimization techniques.

Model parameters

Like in any type of neural network, the NMT model parameters consist of *weight matrices* and *bias vectors*. For a vanilla RNN encoder-decoder model, hidden unit $\boldsymbol{h}_t$ at time step $t \geq 1$ (each rectangle in Figure 2.3) takes two inputs: an input vector $\boldsymbol{x}_t$ (arrows pointing up in Figure 2.3) and a hidden state vector $\boldsymbol{h}_{t-1}$ (arrows pointing right in Figure 2.3). It then multiplies $\boldsymbol{x}_t$ with a weight matrix W_{hx} and $\boldsymbol{h}_{t-1}$ with a weight matrix W_{hh}, adds a bias vector $\boldsymbol{b}_h$, and applies a non-linearity, e.g., a $\tanh$ function, to the resulting vector:

$$\boldsymbol{h}_t = \tanh\left(W_{hx}\boldsymbol{x} + W_{hh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_h\right). \tag{2.10}$$

In the final output layer, we do not compute a hidden state vector, but instead we obtain a linear score vector by multiplying hidden layer input $\boldsymbol{h}_t$ and the attention context vector $\boldsymbol{c}_t$ with weight matrix W_{sh}, before adding bias vector $\boldsymbol{b}_s$:

$$\boldsymbol{s}_t = W_{sh}[\boldsymbol{h}_t; \boldsymbol{c}_t] + \boldsymbol{b}_s. \tag{2.11}$$

In summary, for the vanilla RNN NMT architecture discussed in this section, the following high-dimensional sets of parameters need to be optimized: $W_{hx} \in \mathbb{R}^{|\boldsymbol{h}| \times N}$ ($N = |\boldsymbol{m}|$ in the first hidden layer, $N = |\boldsymbol{h}|$ in the next hidden layers), $W_{hh} \in \mathbb{R}^{|\boldsymbol{h}| \times |\boldsymbol{h}|}$, $W_{sh} \in \mathbb{R}^{|V_\mathcal{E}| \times \boldsymbol{h}}$, $\boldsymbol{b}_h \in \mathbb{R}^{|\boldsymbol{h}|}$, and $\boldsymbol{b}_s \in \mathbb{R}^{|V_\mathcal{E}|}$. Typically, the embedding size $|\boldsymbol{m}|$ and hidden layer size $|\boldsymbol{h}|$ are in range 500–2000, and the vocabulary size $|V_\mathcal{E}|$ is between 30K and 80K. For an LSTM encoder-decoder, similar additional weight matrices exist for the input, output, and forget gates (see Section 2.3.2).

Training process

NMT training operates in two steps, a *forward pass* and a *backpropagation* step (Rumelhart et al., 1986). First, in the forward pass, the model computes its output probabilities. This is done using (2.10) and (2.11), followed by applying a *softmax* normalization function to the output vector $\boldsymbol{s}_t$ to obtain word probabilities over the output vocabulary at time step t:

$$\boldsymbol{p}_t = \text{softmax}\left(\boldsymbol{s}_t\right). \tag{2.12}$$

Next, since during training we know the actual output word e_t, we can compute a *loss function*, which is the negative logarithm of the output probability of e_t:

$$\ell = -\log\left(\boldsymbol{p}_{e_t}\right). \tag{2.13}$$

The second step of NMT training is the *backpropagation* step, which takes the loss ℓ as its input and traverses back through the network, computing gradients with respect to each set of parameters. Optimization of the model parameters—with the goal of reducing the loss—is then done by taking a small step in the direction of the gradients. The size of this step depends on a pre-defined *learning rate* η and on the magnitude of the gradient.

Next, given an updated set of parameter values, we repeat the forward pass and the backpropagation step, until the loss eventually converges to a stable minimum. A common method to perform this iterative process is *stochastic gradient descent* (SGD).

Epochs and mini-batches

Updating the model parameters can be done at different intervals. In the extreme cases, parameter values are updated after every single training sentence pair (i.e., *online* learning) or after observing the entire training bitext (i.e., *batch* learning). Online learning converges quickly but is overly influenced by the most recently observed training examples, while batch learning is slow but more stable. A compromise between these extremes is *mini-batch* learning, in which we calculate gradients after observing n training examples, where typically $n \in [8, 128]$ sentences.

Even with online or mini-batch learning, NMT training requires iterating over the training bitext multiple times before converging to a stable minimum. One pass over the entire training corpus is referred to as a training *epoch*. In order to avoid overfitting and to know when to stop training, it is common to measure after every epoch how well the model performs on an unseen *validation set*. Typically, the set of parameters yielding the highest performance (or lowest *cross-entropy* or *perplexity*) on the validation set is selected to use for translation of unseen test data.

2.3.6 Decoding

Similar to PBMT, decoding of an unseen test set is done using beam search, with the difference that in NMT hypotheses are organized by their actual length rather than by the number of translated source words. Given a beam width b, only the top b distinct hypotheses with the highest probability $p(E \mid F)$ are kept at any time step. It has been shown that beam search decoding tends to favor short output translations, which can be avoided by normalizing the log probability of the sentence under translation for sentence length (Cho et al., 2014a):

$$\hat{E} = \arg\max_{E} \frac{\log p(E \mid F)}{|E|}. \tag{2.14}$$

Recently, a number of open-source NMT systems have been released: Nematus[6] (Sennrich et al., 2017), Neural Monkey,[7] OpenNMT,[8] and seq2seq.[9] All systems have setups similar to the one described in this section but also include many additional

[6] https://github.com/EdinburghNLP/nematus
[7] https://github.com/ufal/neuralmonkey
[8] https://opennmt.net
[9] https://google.github.io/seq2seq/

features. NMT and sequence-to-sequence modeling are explained in more detail in the ACL2016 Neural Machine Translation Tutorial[10] and by Neubig (2017).

2.4 Experimental setup

In this section we describe the general experimental setup used for the experiments described in the research chapters of this thesis.

2.4.1 Oister and general PBMT setup

We run all of our PBMT experiments using *Oister*, an in-house system similar to Moses (Koehn et al., 2007) and developed in *Perl*. Unless stated otherwise in the research chapters, we use the following settings, toolkits and algorithms:

Word alignment of the bitext is performed with GIZA++ (Och and Ney, 2003) in both directions, generating the symmetric alignments using the 'grow-diag-final-and' heuristics. Alignment symmetrization and phrase extraction are done using scripts from the Moses toolkit.

Translation models include bidirectional phrase translation probabilities up to a maximum phrase length of 7, and bidirectional lexical weighting features (Koehn et al., 2003). Phrase and word penalty feature weights are determined automatically during tuning.

Reordering models all use lexicalized reordering, distinguishing between monotone, swap, and discontinuous reordering, with respect to the previous and the next phrase (Tillmann, 2004). We use a distortion limit of 5. The weight of the distortion cost is determined automatically during tuning.

Language models for the target language are linearly interpolated 5-gram models trained with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1999). We use SRILM (Stolcke et al., 2002) to compute our LMs.

Tuning is done using pairwise ranking optimization (PRO, Hopkins and May (2011)). During tuning, 14 parameter estimation runs are performed in parallel on different samples of the n-best list after each decoder iteration. The weights of the individual runs are then averaged and passed on to the next decoding iteration.

Decoding is done with threshold pruning and a beam width of 0.1.

2.4.2 Tardis and general NMT setup

We run all of our NMT experiments using *Tardis*,[11] an in-house system similar to one described by Luong et al. (2015a) and developed in *Torch*. We use the following settings:

[10] `https://sites.google.com/site/acl16nmt/`
[11] `https://github.com/ketranm/tardis`

Word embedding vectors are trained as part of the end-to-end NMT model and have a size of 1,000. We initialize the embedding values by sampling from a Gaussian distribution with unit variance.

Encoder-decoder models are four-layer unidirectional LSTMs with layer sizes of 1,000. No input feeding is used, which comes at the cost of optimal translation quality but allows for a relatively fast realization of large-scale experiments.

Attention is computed using the global attention mechanism with the dot product function as formulated in (2.7) and described by Luong et al. (2015a).

Parameter optimization is done using stochastic gradient descent (SGD), with a mini-batch size of 64 and an initial learning rate of 1, which is decayed by a factor two every epoch after the fifth epoch. All network parameters are uniformly initialized. In addition, we use dropout with probability 0.3 to prevent the network from overfitting (Srivastava et al., 2014).

Decoding is done with a beam size of 12 and using sentence length normalization as described in (2.14).

2.4.3 Evaluation

We evaluate all our translation experiments using the *Bilingual Understudy Evaluation* metric, better known as BLEU (Papineni et al., 2002). BLEU is a precision-based evaluation metric that counts how many n-grams in the translation output of a test set match with n-grams in one or more human reference translations. To prevent BLEU from rewarding short sentences that have a perfect match with words in the references, it also includes a *brevity penalty* that penalizes short translations. Unless specified otherwise, we compute case-insensitive BLEU up to and including n-grams of length 4.

We measure statistically significant differences between two experimental outcomes using approximate randomization (Riezler and Maxwell, 2005), which measures how many times out of 2,000 random permutations we observe a BLEU difference that is similar to or larger than the experimental finding. This number is then divided by 2,000 to obtain the p-value of the experiment. We use $p \leq 0.05$ and $p \leq 0.01$ to indicate weak or strong statistical significance, respectively.

2.5 Summary

In this chapter we have given high-level introductions on two common MT paradigms: phrase-based machine translation (PBMT) and neural machine translation (NMT). The remainder of this thesis consists of six research chapters. In all chapters we run experiments and perform analyses using PBMT. In the final research chapter, Chapter 8, we also use NMT for our experiments. Finally, in Chapter 9 we reflect on future work and what NMT can learn from findings in PBMT and findings in this thesis.

Part I

The Role of Genres in Phrase-Based Machine Translation

The first part of this thesis addresses domain and genre adaptation for PBMT. While domain adaptation has been an active field of research, the definition of a domain is often ambiguous, confusing the concepts *genre*, *topic*, and *provenance*. Most previous work uses in-domain and out-of-domain data that differs at both the genre and the topic level, making it unclear whether the proposed adaptation methods help deal with topic or genre differences.

Despite these unclear definitions, many adaptation approaches depend on the availability of 'domain' (i.e., provenance) information in both the training and the evaluation data. While this dependency is sensible in controlled research scenarios, it also limits the applicability of the proposed methods. For example, in online scenarios the translation task is often of unknown origin, hence an MT system can not in advance be adapted to the task. Even in research settings, provenance labels may not always be available or may be uninformative, e.g., corpus numbers do not necessarily reflect the most useful categorization for a specific translation task.

In this research theme we first demystify the concept of a domain by formulating clear definitions of the two intrinsic text properties genre and topic, before investigating the impact of both aspects on PBMT quality (Chapter 3). Since we find that genre differences pose a larger challenge to PBMT than topic differences, we then propose two methods for genre adaptation. The first method improves translation of a multi-genre test set by (i) enhancing the bilingual training data with genre-specific resources, and (ii) automatically classifying the genre of incoming test documents (Chapter 4). The second method alters an existing weighting approach by relieving its dependency on subcorpus labels in the training data and instead using intrinsic genre-revealing textual features (Chapter 5).

3

Genre and Topic Differences in Phrase-Based Machine Translation

3.1 Introduction and research questions

Training corpora for statistical machine translation (SMT) are typically collected from a wide variety of sources and therefore have varying textual characteristics such as writing style and vocabulary. The test set, on the other hand, is much smaller and usually more homogeneous. The resulting mismatch between the test data and the majority of the training data can lead to suboptimal translation performance. In such situations, it is beneficial to adapt the translation system to the translation task at hand, which is exactly the challenge of *domain adaptation* in SMT.

In domain adaptation research, the test data is considered *in-domain* data, while (the majority of) the training data is referred to as *out-of-domain* or general domain data. However, the concept of a *domain* itself is not unambiguously defined across existing domain adaptation methods. Often, different domains correspond to different subcorpora, in which documents exhibit a particular combination of genre and topic, and optionally other textual characteristics such as dialect and register. This definition has two major shortcomings. First, subcorpus-based domains depend on provenance information, which might not be available, or on manual grouping of documents, which is labor-intensive and often carried out according to arbitrary criteria.

Second, the commonly used notion of domain neglects the fact that *topic* and *genre* are two distinct properties of text (Lee and Myaeng, 2002; Stein and Meyer Zu Eissen, 2006). Two texts can discuss a similar topic, but using very different styles: a news article about Nelson Mandela differs substantially in style from a blog post, tweet, or editorial piece about the same person. Similarly, two news articles, one about sports and the other about politics, share the same genre but have very different vocabulary. Since most previous work on domain adaptation for SMT uses in-domain and out-of-domain data that differ on both the topic and the genre level, it is unclear whether the proposed solutions help deal with topic or genre differences.

In this chapter we take a step back and first disentangle the concepts topic and genre, then we analyze and quantify their effect on PBMT, which we believe is a necessary step towards further improving domain adaptation. Concretely, we ask:

RQ1 *What impact do genre and topic differences have on PBMT quality?*

To answer this question, we first need clear definitions of the different aspects that together make up a domain. To this end, we ask:

RQ1a. *Can we clarify the ambiguous use of the concept* domain *with regard to adaptation in PBMT?*

After providing clear definitions of the concepts topic and genre, we introduce a new annotated benchmark set that has controlled topic and genre distributions. By translating our new benchmark set with a genre-balanced PBMT system and measuring the respective translation quality of topics and genres, we ask:

RQ1b. *Which of two intrinsic text properties, topic and genre, presents a larger challenge to PBMT?*

Next, we perform a number of quantitative and qualitative analyses to get a deeper understanding of the differences among topics and genres in the context of PBMT. First, we introduce new quantitative metrics that compute how well each genre or topic is covered by the PBMT models. Not only do we evaluate the portion of unknown source *words*, but we also compute model coverage of source *phrases* and source-target *phrase pairs*. Together with a manual annotation of observed out-of-vocabulary (OOV) words, these measures enable us to answer the question:

RQ1c. *To what extent do topic and genre differ with respect to out-of-vocabulary words and phrases?*

Organization. This chapter is organized as follows: In Section 3.2 we provide an overview of existing work on domain adaptation for PBMT. Next, in Section 3.3 we disentangle the concept *domain* by providing clear definitions of *topic* and *genre*, and we introduce our new benchmark set with controlled genre and topic distributions, the Gen&Topic data set. In Section 3.4 we use this new benchmark to quantify and analyze the impact of different genres and topic on PBMT. Finally, we discuss conclusions and implications of this work in Section 3.5.

3.2 Domain adaptation for PBMT

Domain adaptation is the task of adjusting or modifying a statistical model learned from some type of data (the *source* domain) in such a way that it is useful for another type of data with a different distribution (the *target* domain) (Ben-David et al., 2010). Typically, labeled data for the target domain, i.e., in-domain data, is scarce, yielding models that are adapted to the target domain but have weak statistical power. On the other hand, labeled data for the source domain, i.e., out-of-domain data, is usually plentiful. It is therefore often beneficial to exploit (at least part of) the out-of-domain data in addition to the available in-domain data.

In the scenario of PBMT, where domain adaptation was first addressed by Langlais (2002), labeled data consists of parallel corpora, which are only abundant for a few

Table 3.1: Categorization of domain adaptation strategies for PBMT. Detailed descriptions of each strategy are provided in the text, in the same order as the listing in this table.

Level of adaptation	Adaptation strategy	Cross-domain or dynamic adaptation?	Applicable to TM or LM?
Data level	Data selection	Cross-domain	TM, LM
	Data augmentation	Cross-domain	TM, LM
	Data clustering	Cross-domain, dynamic	TM, LM
Model level	Mixture modeling	Cross-domain, dynamic	TM, LM
	Phrase-table fill-up	Cross-domain	TM
	Feature augmentation	Cross-domain	TM, LM
	Instance weighting	Cross-domain, dynamic	TM
	Word alignment adaptation	Cross-domain	TM
	Domain-specific classifiers	Dynamic	TM, LM

text types such as news articles and parliamentary proceedings. For other types of text, parallel translations can be hard to obtain, and the resulting PBMT models are usually weak. Using large amounts of out-of-domain data, however, is also not optimal due to differences in writing style and vocabulary between the training and test data.

To address these problems, domain adaptation for PBMT has been an active research area in recent years. Since this thesis builds upon previous findings and achievements, we provide in this section an extensive overview of related research in domain adaptation. To this end, different categorizations of domain adaptation methods can be made. In the following we group approaches based on whether the main adaptation process takes place at the data or the model level. However we also distinguish (i) whether the target domain is known in advance, i.e., *cross-domain* adaptation, or adaptation is performed dynamically at translation time, i.e., *dynamic* adaptation, and (ii) whether adaptation applies to the language model (LM) or the translation model (TM). Table 3.1 provides a summary of the main types of adaptation methods.

3.2.1 Domain adaptation at the data level

A straightforward way of adapting a PBMT system to a new domain is by applying *data selection*. The goal of data selection is to either select or discard training data based on their relevance for the domain of interest. Data selection can be applied monolingually for LM adaptation (Gao et al., 2002; Moore and Lewis, 2010) or bilingually for TM adaptation (Yasuda et al., 2008; Axelrod et al., 2011; Duh et al., 2013; Cuong and Sima'an, 2014a; Axelrod et al., 2015; Chen and Huang, 2016). While data selection is typically performed in a cross-domain adaptation fashion, Hildebrand et al. (2005) vary the selection size and composition between different test sentences, however this is not feasible at translation run time since new models have to be trained for each selection. We further discuss and explore data selection methods in Chapter 8.

Instead of removing irrelevant training data, one can also include additional relevant training data to adapt a PBMT system. This strategy is applied in *data augmentation*

methods, which aim at obtaining or synthesizing parallel in-domain data. In order to acquire additional bilingual training data, some approaches extract data from in-domain comparable corpora (Munteanu and Marcu, 2005; Daumé III and Jagarlamudi, 2011; Irvine et al., 2013b), while others opt for domain-focused web-crawling (Pecina et al., 2011, 2015). A second data augmentation procedure involves the generation of synthetic parallel data from in-domain parallel data. These methods use a pre-trained out-of-domain PBMT system to translate in-domain monolingual data which is then used as in-domain bilingual training data. Translations can be synthesized from the source to the target language (Ueffing, 2006; Schwenk, 2008), or from the target to the source language. The latter has been shown to yield the largest improvements (Bertoldi and Federico, 2009).

Most of the above approaches rely on the assumption that domain information is available for the complete training corpus. This is, however, not always a realistic scenario, especially when data is harvested from the web. *Data clustering* methods circumvent this shortcoming by performing automatic clustering of training sentences (Yamamoto and Sumita, 2008; Sennrich, 2012a; Carpuat et al., 2014) or sentences in the development set (Sennrich et al., 2013), such that a single system can be adapted to multiple domains.

3.2.2 Domain adaptation at the model level

Given a certain amount of in-domain (or 'pseudo in-domain' if obtained with selection or augmentation) data, there are several techniques to use this data in the PBMT models. A common strategy to is *mixture modeling* (Friedman et al., 2001), in which two or more TMs or LMs are interpolated using weights representing their respective relevance. The available in-domain data is used to create PBMT models that are tailored to the given translation task but have poor coverage, while the out-of-domain data serves to train large PBMT models with good coverage but low similarity to the domain of interest. While it is often preferable to use small, specific models rather than large, general models, even better results can be obtained by interpolating both types of models, either log-linearly (Koehn and Schroeder, 2007) or linearly (Foster and Kuhn, 2007). The latter also investigate cross-domain versus dynamic model interpolation and conclude that for cross-domain adaptation both TM and LM interpolation work well, while for dynamic adaptation only LM interpolation is profitable. Sennrich (2012b) shows that mixture modeling can be extended to multiple (4 or 10) translation models by learning interpolation weights through translation model perplexity minimization.

Another strategy to combine different models is to use *domain-specific classifiers* that route an incoming sentence or document to pre-trained domain-specific PBMT systems (Xu et al., 2007; Banerjee et al., 2010; Song et al., 2011; Wang et al., 2012). Since such classifiers can make different decisions for different sentences at decoding time, this is a dynamic adaptation approach. We further discuss and explore this strategy in Chapter 4.

Next, an alternative model combination approach is the *fill-up technique* (Bisazza et al., 2011), a method that starts with an in-domain phrase table (or reordering model), and extends this with entries from a larger out-of-domain model. An additional feature indicates whether or not each phrase pair is observed in the in-domain model. The

benefits of this strategy are confirmed by Haddow and Koehn (2012), who analyze the effects of adding in-domain data at various training stages, and conclude that the best results are obtained when using out-of-domain data to improve coverage of rare words.

Related to this are *feature augmentation* methods, which extend the PBMT models with additional features, for example by creating domain-specific copies of the original log-linear feature set, such that adaptation is done automatically during system optimization (Daumé III, 2007; Clark et al., 2012). Cuong et al. (2016) include features indicating to what extent phrases are domain-specific or phrase pairs are coherent across domains. Feature augmentation methods have the advantage of being extendable to multiple domains, albeit at the cost of increased computational complexity.

A large body of work has performed feature augmentation based on the relevance of specific training data. Such *instance weighting* methods prioritize bilingual training instances that are most relevant to the development and test data, by assigning weights to sentence pairs (Matsoukas et al., 2009) or phrase pairs (Foster et al., 2010; Chen et al., 2013; Cuong and Sima'an, 2014b). Dynamic variants of this technique typically use a vector space model (VSM) in which each training phrase pair is represented by a vector. This vector captures for example the phrase pair's source context (Costa-jussà and Banchs, 2011) or latent representations learned with Latent Semantic Indexing (LSI, Landauer et al. (1998)) (Banchs and Costa-Jussà, 2011). At test time, the cosine distance between the test sentence under translation and each phrase-pair vector is added as a feature indicating the phrase pair's relevance. Instead of computing a similarity score, Chen et al. (2014) add vector entries based on subcorpus frequencies as direct features to the phrase table. We further discuss and explore adaptation by means of instance weighting in Chapter 5.

Finally, *word alignment adaptation* methods aim at improving domain-specific word alignments, for example using general-domain and domain-specific translation dictionaries (Wu and Wang, 2004), by interpolating different word alignment models (Wu et al., 2005; Civera and Juan, 2007), or by inducing latent sub-domains in a heterogeneous bilingual corpus (Cuong and Sima'an, 2015). Duh et al. (2010) show that adding data during word alignment but removing it before phrase extraction can help decrease lexical translation ambiguity.

3.3 Genre and topic differences in PBMT

The definition of a domain varies across the above described previous work on domain adaptation and is often imprecise. Typically, domain means 'different data set,' and is therefore a hard-labeled concept that corresponds to provenance or particular topic-genre combinations. In this chapter, we explicitly note that *topic* and *genre* are two distinct properties of text (Lee and Myaeng, 2002; Stein and Meyer Zu Eissen, 2006). Both concepts have been mentioned in previous work on domain adaptation, but definitions are often ambiguous and overlapping. While some work illustrates the challenge of domain adaptation using vocabulary difference examples, others note that different domains are characterized by variations in writing style. A few previous efforts have analyzed the impact of domain adaptation on PBMT (Duh et al., 2010; Haddow and Koehn, 2012; Irvine et al., 2013a), however these all use subcorpora as domains and

do not provide any insight in how topic and genre differences affect PBMT. Based on previous work, it is unclear whether proposed solutions help deal with topic or genre differences.

3.3.1 Definitions

To allow for a more focussed discussion of the impact of topic and genre on PBMT, we avoid in this chapter using the ambiguous term *domain*, and instead focus on the text properties *topic*, *genre* and *provenance*:

Topic is the general subject of a document. Topics can be determined on multiple levels, ranging from very broad to more detailed. Examples of topics include sports, politics, and science, but one could also opt for more fine-grained topics like football, tennis, and swimming.

Topics are typically associated with vocabulary and word sense differences, for example illustrated by Hasler (2014) with the French word 'noyau', which translates as 'nucleus' in scientific topics, as 'core' in financial topics, and as 'kernel' in technical topics.

Genre is harder to define, as there is no single definition in literature. For example, Swales (1990) refers to it as a class of communicative events where there is some shared set of communicative purposes, while Karlgren (2004) calls it a grouping of documents that are stylistically consistent and intuitive to accomplished readers of the communication channel in question. Based on previous definitions, Santini (2004) concludes that the term genre is used as a concept complementary to topic, covering the non-topical text properties function, style, and text type. Like topics, genres can also exhibit different levels of granularity (Lee, 2001). Examples of genres include formal or informal text (high-level), and newswire, editorials, and user-generated text (low-level).

Genres are typically characterized by stylistic aspects such as writing style, distributions of pronouns and stopwords, and register. For example, informal genres are more likely to exhibit a casual register, reflected by the use of contractions (*ain't*) or acronyms (*LOL* meaning *laughing out loud*).

Provenance indicates the origin of a document, typically expressed by the document's collection or subcorpus name. In previous work, provenance is sometimes accompanied by manually assigned genre labels, however these may not be available. In this thesis we use the term provenance and subcorpus interchangeably.

Although in theory any topic and genre can co-occur, some combinations are more common than others, e.g., technical manuals are more likely to be about smartphones than about politics. While both topic and genre are intrinsic properties of text, provenance information has to be provided explicitly to be of use for adaptation in PBMT. Still, most previous work on domain adaptation uses exactly this type of information to adapt PBMT systems to a specific translation task, see Section 3.2. In this chapter we disregard provenance information as much as possible, as it is not guaranteed to be available, and focus on topics and genres.

3.3.2 Genre and topic adaptation

Topic adaptation for PBMT has been explicitly addressed in recent years (Tam et al., 2007; Eidelman et al., 2012; Hasler et al., 2012, 2014a,b,c; Hewavitharana et al., 2013; Su et al., 2015), mostly using topic modeling. These methods assume that in a document collection, i.e., a parallel or monolingual corpus, a number of latent topics exist, which are all represented by their most common words. At training time, these topics are detected using different flavors of latent semantic analysis (LSA, Deerwester et al. (1990)) or latent Dirichlet allocation (LDA, Blei et al. (2003)), and the resulting topic probabilities are added as features to the translation model or the language model (Ruiz and Federico, 2011, 2012). At decoding time, a distribution over the considered latent topics is inferred for each test document or sentence. Based on this distribution, adaptation takes place dynamically at the document or sentence level.

Genre adaptation for PBMT has not been studied as extensively as topic adaptation, and approaches dubbing their contribution 'genre adaptation' usually adapt to the same subcorpora that are used for domain adaptation. One piece of previous work has explicitly addressed style adaptation of the language model (Bisazza and Federico, 2012). However, this method requires the availability of clearly separable training corpora, and thus also really adapts to provenance rather than genre.

3.3.3 The Gen&Topic benchmark set

To analyze the impact of genre and topic differences in PBMT, we need a test set where both dimensions are controlled as much as possible. Unfortunately, currently available and commonly used benchmarks meet this requirement only to a limited degree. For instance, while the NIST OpenMT sets do contain documents drawn from two genres, newswire and web, both genres exhibit a different distribution over topics, i.e., the same topic might not be equally represented across genres, and vice versa.

To overcome this limitation, we introduce a new Arabic-English parallel benchmark, the Gen&Topic data set, that contains documents with controlled topic and genre distributions. This benchmark consists of manually translated news articles crawled from the web with their corresponding, manually translated readers' comments, and thus comprises the genres *newswire* (NW) and *user-generated* (UG) text. Since each pair of NW and UG documents originates from the same article, we assume that both documents discuss the same topic, for which labels are provided by the source websites. By including comparable numbers of tokens per genre for each article, we enforce equal topic distributions across genres. Examples of NW-UG pairs are shown in Table 3.2. Note that the UG sentences in the Gen&Topic data set are well-formulated comments rather than dialogue-oriented content such as SMS or chat messages. Further details on the web crawling efforts from which the Gen&Topic data set has emerged are given in Section 4.3.

For parameter optimization purposes, we split the complete benchmark into a development and a test set, such that the development set contains approximately one-third of the data, while ensuring that articles in each set originate from non-overlapping time periods. Table 3.3 lists the specifications of the complete benchmark, which is available for download at `http://ilps.science.uva.nl/resources/gen-topic/`.

Table 3.2: English-side samples from the Gen&Topic data set. All pairs of newswire (NW) and user-generated (UG) fragments in the data set discuss the same article and are topically related.

Topic	Newswire sentence	User-generated sentence(s)
Culture	The 12 contestants competed during a May 3rd Prime before a panel of judges and millions of viewers across the Arab world.	Your program's name is "Arab Idol", which is in English, and you allowed Barwas to participate and represent Iraq while she sings in Kurdish!!!
Economy	Yemen is mulling the establishment of 13 industrial zones across its six planned administrative regions in a bid to stimulate development and create job opportunities.	What development in Yemen are you talking about? We will continue to call for freedom until independence and liberation and the routing of the northern occupation from our lands.
Health	The availability of a defibrillator at the American University sports club saved the life of 21-year-old student Munir Khalil.	May Allah give you well-being! May Allaah have mercy upon him. This machine is in the USA and the government places it in the public places.
Politics	Eager to satisfy public aspirations for an economic recovery, presidential candidates are proposing a mix of mega projects and smaller or non-traditional ones, experts say.	There are questions that are without answers regarding Egypt since the black historical date of 1952?!! Isn't 60 years of lies and failings enough for you?
Security	Some cafe owners have taken precautionary measures to protect their establishments and to ensure the safety of their customers.	Which group commits them and what is their purpose? Poetry is the noble Iraqi's main nutrient; nothing can prevent it, not bombing or anything else.

3.4 The impact of genre and topic differences on PBMT

To quantify the impact of multiple genres and topics in a test corpus, we run a series of experiments in which we measure translation quality, model coverage, and observed out-of-vocabulary (OOV) types.

3.4.1 Translation quality

We first run a translation experiment on the Gen&Topic test set using our in-house PBMT system Oister (see Section 2.4.1). The 5-gram linearly interpolated language model covers all topics and genres contained in the benchmark. We tune our system on the complete Gen&Topic development set, covering both genres and all topics.

Naturally, performance differences across topics and genres depend on the degree to which both are represented in the parallel training data. To allow for a fair comparison, we down-sample our available training data to be as balanced as possible in terms of topics and genres. The resulting system is trained on approximately 200K sentence pairs with 6M source tokens per genre, see Table 3.4, as much as is available for UG. All data originates from the same web sources as the documents in the benchmark. Our

Table 3.3: Statistics of the Arabic-English Gen&Topic data set containing five topics and two genres: newswire (NW) and user-generated (UG) text. Tokens are counted on the Arabic side.

Topic		Genre		Total
		NW	UG	
Culture	Lines	654	507	1161
	Tokens	15.5K	14.9K	30.4K
Economy	Lines	500	578	1078
	Tokens	16.0K	15.5K	31.5K
Health	Lines	384	319	703
	Tokens	9.7K	9.3K	19.1K
Politics	Lines	494	646	1140
	Tokens	15.8K	15.8K	31.6K
Security	Lines	532	826	1358
	Tokens	16.1K	15.9K	32.0K
Total	Lines	2564	2876	5440
	Tokens	73.2K	71.3K	144.5K

more competitive system (see Chapter 6) that uses also LDC-distributed data achieves slightly better performance, but is more favorable for NW than for UG translation tasks. Due to the strict data requirements in terms of topic and genre distributions, as well as the availability of sizable parallel training data, our current experimental set-up covers Arabic-English only. We tokenize all Arabic data using MADA (Habash and Rambow, 2005).

Table 3.4: Arabic-to-English down-sampled training data specifications of our balanced genre system. Tokens are counted on the English side.

	Lines	Tokens
Newswire	206K	6.2M
User-generated	189K	6.2M
Total	395K	12.4M

Table 3.5 compares BLEU scores of the Gen&Topic data, split down by topics and genres. We observe that translation performance fluctuates much more across genres than across topics: there is a large gap of 3.9 BLEU points between NW and UG, which can be entirely attributed to actual genre differences given the construction of the Gen&Topic data set and the use of down-sampled training data. On the other hand, the gap between different topics is only 0.6 BLEU points on average, and at most 1.1 (between culture and politics). A translation quality gap between genres has also been observed in past OpenMT evaluation campaigns. However, as the NIST benchmarks have not been controlled for topics across genres, it is unclear to what extent this gap can be attributed to actual genre differences.

Table 3.5: Arabic-to-English BLEU scores on the Gen&Topic test set (1 reference translation) per topic-genre combination. Tuning was done on the complete Gen&Topic development set. Variations in translation quality are represented by average pairwise BLEU score differences.

	NW	UG	All	
Culture	19.2	17.6	19.3	
Economy	19.9	15.9	18.9	
Health	19.3	17.7	18.8	Avg. diff.: ±0.6
Politics	21.3	13.6	18.2	
Security	19.3	16.2	18.5	
All	19.9	16.0	18.9	
	Avg. diff.: ±3.9			

3.4.2 Model coverage analysis

Next, to explain the large performance gap between genres, we analyze the phrase lengths within Viterbi translations, source phrase and phrase pair recall, and phrase pair OOV of the Gen&Topic test set. All quantitative results of these analyses are shown in Table 3.6.

Average source-side phrase length

We first compute the average number of source words contained in the phrases that our PBMT system uses to produce the 1-best translations for the Gen&Topic test set. One can see that UG is translated with shorter phrases than NW, and that differences are more pronounced between genres than among topics. This difference, in turn, can be due to unreliable translation probabilities but also to the mere lack of translation options in the models. We quantify the impact of the latter by measuring phrase recall on each test portion.

Phrase recall and phrase pair OOV

To compute phrase recall, we first automatically word-align the test set and extract from it a set of reference phrase pairs using the same procedure applied to the training data. Then, we count the number of reference phrase pairs whose source side is covered by the translation models (*source phrase recall*) and the number of reference phrase pairs that are fully covered by the translation models (*source-target phrase pair recall*). Formally, we define the set of source-matching phrases as:

$$M^S = \{(\bar{f}, \bar{e}) \mid (\bar{f}, \bullet) \in P_{test} \wedge (\bar{f}, \bullet) \in P_{train}\}, \tag{3.1}$$

where P_d refers to the set of phrase pairs $(\bar{f}, \bar{e})$ that can be extracted from corpus d. Source phrase recall R_n^S for phrases of length n is then defined as:

Table 3.6: Impact of genre and topic differences on various indicators of PBMT model quality.

Gen&Topic portion	BLEU	Avg.phr. length	Source phrase recall 1	2	3	4+	Src-trg phrase pair recall 1	2	3	4+	Phr.pair OOV
NW	19.9	1.45	99.3	81.4	41.8	7.1	73.8	39.4	13.7	1.8	71.5
UG	16.0	1.38	97.2	74.7	36.0	6.3	56.2	28.8	8.7	1.1	76.0
Culture	19.3	1.39	98.2	77.6	36.5	5.3	66.2	35.2	10.7	1.2	74.2
Economy	18.9	1.42	98.4	78.7	39.4	6.5	65.3	33.5	10.9	1.4	73.8
Health	18.8	1.41	98.3	76.6	37.1	5.4	64.5	33.5	11.0	1.2	75.2
Politics	18.2	1.41	98.1	78.6	39.8	7.7	60.8	33.1	11.2	1.5	73.4
Security	18.4	1.42	97.6	77.0	40.2	8.4	62.7	33.3	11.6	1.8	73.3

$$R_n^S = \frac{\sum_{(\bar{f},\bar{e}) \in M^S \wedge |\bar{f}|=n} c_{test}\left(\bar{f},\bar{e}\right)}{\sum_{(\bar{f},\bar{e}) \in P_{test} \wedge |\bar{f}|=n} c_{test}\left(\bar{f},\bar{e}\right)}, \tag{3.2}$$

where $c_{test}\left(\bar{f},\bar{e}\right)$ denotes the frequency of phrase pair $\left(\bar{f},\bar{e}\right)$ in the test set. Analogously, we define the set of source-target-matching phrase pairs as:

$$M^{S,T} = \{\left(\bar{f},\bar{e}\right) \mid \left(\bar{f},\bar{e}\right) \in P_{test} \wedge \left(\bar{f},\bar{e}\right) \in P_{train}\}, \tag{3.3}$$

and the source-target phrase pair recall $R_n^{S,T}$ for phrases of length n as:

$$R_n^{S,T} = \frac{\sum_{(\bar{f},\bar{e}) \in M^{S,T} \wedge |\bar{f}|=n} c_{test}\left(\bar{f},\bar{e}\right)}{\sum_{(\bar{f},\bar{e}) \in P_{test} \wedge |\bar{f}|=n} c_{test}\left(\bar{f},\bar{e}\right)}. \tag{3.4}$$

Finally, we call *phrase pair OOV* the portion of reference phrase pairs that are not covered by the translation models, that is: $1 - \sum_n^N R_n^{S,T}$, where N is the phrase limit used for phrase extraction.

The results of our analysis, broken down by source phrase length, show that source phrase recall is much lower in UG than in NW, while variations among topics are only very small. The stronger impact of genre differences is even more visible on phrase pair recall: for instance, our system knows the correct translation of 73.8% of the single-source-word phrase pairs in the NW genre. In UG this is only 56.2%, despite the equal amounts of training data per genre in our system. These figures suggest that model coverage—both mono- and bilingual—is an important reason for the low translation quality on UG data.

Most existing approaches to domain adaptation focus on domain-sensitive scoring or selection of existing translation candidates (Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011; Chen et al., 2013, among others). This strategy is supported by the error analysis of Irvine et al. (2013a), who show that scoring errors are more common across domains than errors caused by OOVs, in the source as well as the target language. Across genres however, our results in Table 3.6 show that both word-level and

Table 3.7: Categorization of (sub)classes used for our manual OOV analysis.

Main OOV class		Subclasses
Rare:	Rare but correct	Numbers, proper nouns, technical terms, foreign words
Dial:	Dialectal forms	Egyptian future tense, unusual spelling of dialectal form
Morph:	Morphological variants	Unseen morphological variant
Spell:	Spelling errors	Replaced letter, missing or inserted blank
Coll:	Colloquialisms	Unconventional spelling, emphasis with repeated letters

Table 3.8: Examples of OOVs observed in the Gen&Topic set with their respective explanation and main OOV class. See Table 3.7 for details of the OOV classes.

Arabic OOV	English translation	Explanation of OOV	Main OOV class
داعش	ISIL	New proper noun	Rare
هينسوا	(they) will forget	Dialectal future tense	Dial
يقدسون	(they) revere	Third person plural present tense	Morph
توفيرالوظائف	creationofjobs	Missing blank	Spell
المتطوعيين	volunteeeers	Deliberate wrong spelling	Coll

phrase-level OOVs are a more likely explanation for the performance differences. This stresses the need to not only improve model scoring, but to also address model coverage, for example by paraphrasing (Callison-Burch et al., 2006), translation synthesis (Irvine and Callison-Burch, 2014), source sentence simplification (Hasler et al., 2016), text normalization (Bertoldi et al., 2010), or automatic collection of additional sizeable bilingual resources (see Chapter 4). Interestingly, similar conclusions have been drawn for adaptation of part-of-speech (POS) taggers (Plank et al., 2014).

3.4.3 Manual OOV analysis

To get a better understanding of the OOVs observed for the genres and topics in the Gen&Topic set, we perform a fine-grained analysis, for which we manually annotate 500 sentences on the source side (equally distributed over genres and topics) to identify the class of each OOV. Annotations are done for top and sub-level classes, e.g., replaced letter, which is a subclass of spelling errors. In total, we consider 17 subclasses which we group into five main classes, see Table 3.7 for an overview of the considered categories and Table 3.8 for examples.

Table 3.9 shows the *type* level percentages for each main OOV class per genre or topic. When comparing the two *genres*, a number of observations emerge. Firstly, rare but correct words (e.g., proper nouns and technical terms, both regular issues for adaptation in PBMT) make up the vast majority of the OOVs in NW, but are relatively infrequent in UG. By contrast, OOVs containing unseen morphological variants are equally common in both genres. Although complex morphology is language-specific, a

Table 3.9: Type-level error percentages per Gen&Topic portion of main OOV classes, see Table 3.7 for details of the OOV classes. Other events include words that are not understandable or occur in the phrase table but only captured in a different context.

Gen&Topic portion	OOV type					
	Rare	Dial	Morph	Spell	Coll	Other
NW	77.8	0.0	16.7	5.6	0.0	0.0
UG	9.8	9.0	17.2	42.6	12.3	9.0
Culture	17.4	0.0	17.4	52.2	8.7	4.3
Economy	13.8	0.0	34.5	31.0	13.8	6.9
Health	15.8	10.5	15.8	36.8	10.5	10.5
Politics	25.0	25.0	12.5	25.0	0.0	12.5
Security	23.5	8.8	5.9	41.2	14.7	5.9

Table 3.10: Token-level error percentages per Gen&Topic portion of main OOV classes, see Table 3.7 for details of the OOV classes. Other events include words that are not understandable or occur in the phrase table but only captured in a different context.

Gen&Topic portion	OOV type					
	Rare	Dial	Morph	Spell	Coll	Other
NW	89.5	0.0	7.9	2.6	0.0	0.0
UG	11.7	8.6	17.2	41.4	12.5	8.6
Culture	16.7	0.0	20.8	50.0	8.3	4.2
Economy	18.8	0.0	31.3	28.1	15.6	6.3
Health	23.8	9.5	14.3	33.3	9.5	9.5
Politics	25.0	25.0	12.5	25.0	0.0	12.5
Security	48.1	5.8	3.8	28.8	9.6	3.8

rare morphological word in Arabic often maps to a rare multi-word phrase in English, resulting in phrase-level OOVs. Next, not entirely surprising, the majority of OOVs in UG are due to spelling errors. Finally, OOVs assigned to the remaining classes are never observed in NW but occasionally occur in UG.

Next, a comparison of the main OOV classes among the various *topics* shows a few notable distributions. Dialectal forms, for example, are rare in all topics except politics, where they are commonly observed in the form of Egyptian future tense. This can be explained by the presence of news articles about elections in Egypt in the Gen&Topic set. Next, while spelling errors are common in all topics, its abundance is most prominent in culture. Most spelling errors concern missing or inserted blanks, suggesting that comments may have been written on mobile devices. Finally, unseen morphological variants are more frequent in economy than in other topics, however with no conclusive explanation.

Finally, Table 3.10 shows the *token* level percentages for each main OOV class per genre or topic. Patterns are very similar to those observed for word types, with the notable exception that a small number of unseen proper nouns occur repeatedly in a few topics. For example, داعش meaning 'ISIL' is the main topic of a number of security documents, yielding a very high token-level OOV rate for this topic.

3.5 Conclusions

Despite the fact that domain adaptation is an active field of research in SMT, there is little consensus on what exactly constitutes a domain. Typically, different domains correspond to different subcorpora or data sets. Since most previous work on domain adaptation in SMT uses in-domain and out-of-domain data that differ on both the topic and the genre level, it is unclear whether the proposed solutions help deal with topic or genre differences. In this chapter we have disentangled these two concepts and asked:

RQ1 *What impact do genre and topic differences have on PBMT quality?*

To answer this question, we first provided clear definitions of the different aspects that together make up a domain: while *topic* refers to the subject of a text and is mostly characterized by specific vocabulary or word senses, *genre* refers to non-topical properties function, style and text type and is characterized by aspects such as writing style, stopwords, and register.

Using these definitions, we introduced a new annotated Arabic-English benchmark set with controlled topic and genre distributions covering newswire (NW) and user-generated (UG) documents in five topics. When translating this data using a genre-balanced PBMT system, we found that translation performance in BLEU fluctuates much more between the two genres than across topics, and that UG is translated with shorter phrases than NW. Next, we introduced new quantitative metrics measuring *source phrase recall* and *source-target phrase pair recall*, representing the source and source-target coverage of our models. We found that poor model coverage—both mono- and bilingual—is an important reason for the low PBMT quality on UG data, stressing the need for methods improving the PBMT quality of informal, UG data. Finally, our fine-grained manual error analysis at the word level also suggests that source coverage could benefit from text normalization or increasing the amount of relevant training data.

4

Genre Adaptation Using Automatic Classifiers

4.1 Introduction and research questions

Motivated by the observation that genre differences pose a bigger challenge to PBMT than topic differences (see Chapter 3), we further explore the impact of different genres on PBMT quality. While the systematically constructed Gen&Topic data set is limited to the Arabic-English language pair and covers only two genres, we extend our research in this chapter to four genres (colloquial, editorial, news, and speech) and four language pairs (Arabic-English, Bulgarian-English, Chinese-English, and Persian-English), for which we automatically harvest data from the web. Using this diverse setting, we address the following question:

RQ2 *Is the observed impact of genre differences on PBMT consistent among various language pairs and data settings?*

We first look at different language pairs. We study whether and to what extent the observed differences between genres are language-pair specific or stable across language pairs. Concretely, we ask:

RQ2a. *To what extent does the impact of genre differences on PBMT vary among language pairs?*

Since translation quality differences between genres likely depend on the amount of available training data covering the genre of interest, we also ask:

RQ2b. *To what extent can differences in PBMT performance among genres be explained by the proportion of genre-specific resources?*

To answer this question, we automatically harvest parallel data from the web covering all four genres of interest, thus addressing the problem of poor coverage we observed in Chapter 3. In addition, we exploit our newly collected corpora to take a first step towards adapting PBMT systems to different genres, which is the central theme of RQ3:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

Most existing adaptation approaches depend on the availability of (manually assigned) corpus or provenance labels and make two strong assumptions, which contribute to the success of these methods: first, the translation task has known domain, genre or topic labels that are exploited to adapt the system. Second, an in-domain development set, that is similar to the test set in terms of writing style and vocabulary, is available to optimize the adaptation approach for a specific translation task. While these are fair assumptions in a controlled research setting, they are less realistic in for example an online MT service. In this chapter, we consider the scenario in which a PBMT system is provided with a test document of unknown origin. We investigate whether we can exploit automatic genre classification to guide each test document to the most appropriate pre-trained system, by asking:

RQ3a. *Can we successfully adapt PBMT systems using automatic genre classifiers?*

Concretely, we use our harvested corpora in four genres to build genre-specific PBMT systems. We then learn accurate document-level genre classifiers, which we use to determine which of the genre-specific systems to employ for the particular document.

Organization. This chapter is organized as follows: After surveying related work in Section 4.2, we describe in Section 4.3 the collection and creation of training and evaluation corpora of four genres in four language pairs. In Section 4.4 we describe how we train baseline PBMT systems for all genres, and we analyze the relation between translation quality and the amount of genre-specific resources. To adapt PBMT to a mixture-of-genres test set, we train automatic genre classifiers in Section 4.5, and use these for genre adaptation in Section 4.6. Finally, we provide conclusions of this chapter in Section 4.7.

4.2 Related work

Three research areas are relevant for this chapter: first, we discuss related work on using web data as parallel corpora. Next, we discuss research on genre classification, which is relevant for our work on training automatic genre classifiers. Finally, we discuss previous work performing adaptation using automatic classifiers.

4.2.1 Web as corpus

The rapid growth of the Internet has resulted in a large body of work on extracting and using web-crawled data for various natural language processing (NLP) tasks, which is reflected by the existence of a workshop series and a special interest group on Web as Corpus (WAC).[1] Besides detecting relevant web pages (Qi and Davison, 2009) and extracting clean textual content (Spousta et al., 2008), data harvesting for MT faces

[1] www.sigwac.org.uk

another challenge: multilinguality. The most useful and interesting web-crawled corpora for MT are bilingual. Preferably, such corpora contain parallel translations; however previous work has shown that the use of *comparable* corpora, from which near-parallel sentence pairs can be extracted, also helps improving PBMT quality (Munteanu and Marcu, 2005, 2006; Tillmann, 2009; Tillmann and Xu, 2009).

While a fair number of approaches exist for web-crawling of parallel corpora, most related to our work are previous efforts to harvest domain- or genre-specific bilingual data. For example, Pecina et al. (2011, 2012) construct a pipeline combining various open-source tools for web crawling, preprocessing, and sentence alignment, and use the resulting corpora for adaptation to two domains. Further experiments also show that parallel development data does not have to be very clean since in-domain tuning is robust to noise (Pecina et al., 2015).

A different line of work focuses on detecting and extracting comparable sentences from microblogs, e.g., Twitter or Sina Weibo, a noisy user-generated genre difficult to automatically translate. To this end, Jehl et al. (2012) use an information retrieval (IR) approach to detect comparable tweets related to the topic "Arab Spring". Carter (2012, Chapter 5) uses a similar IR-based approach as well as a self-learning approach in which Twitter translations are generated by a general-purpose PBMT system, which outperforms the IR approach. This work is not restricted to specific topics, and even abandons the assumption that the test data is known at training time. Finally, Ling et al. (2013b) exploit the fact that people regularly use multiple languages in a single microblog message. Assuming that such messages contain internal parallel segments, they use a dynamic programming approach to detect these parallel parts. All three approaches show that adding web-crawled microblog data to a general-domain training corpus improves translation of microblog test data.

4.2.2 Text genre classification

While traditionally most text classification work focused on classification by topic (Sebastiani, 2002), the past decades have shown an increasing interest in genre classification, which has resulted in various methods that combine different classifiers with different sets of genre-specific features.

Karlgren and Cutting (1994) are among the first to use simple document statistics, such as common word frequencies, first-person pronoun count, and average sentence length. Kessler et al. (1997) use similar features and categorize four types of genre-revealing cues: *structural cues* (e.g., part-of-speech (POS) tag counts), *lexical cues* (specific words), *character-level cues* (e.g., punctuation marks), and *derivative cues* (ratios and variation measures based on other types of cues). Stamatatos et al. (2000) use the most frequent words of the entire written English language together with punctuation mark frequencies to distinguish four text genres. With a discriminant analysis, they determine that the optimal number of most frequent words is 30. Dewdney et al. (2001) compare a large number of presentation features (representing a document's stylistic information) and show that these outperform bag-of-words approaches, which are traditionally used in topic-based text classification, and that the optimal feature set contains both word and presentation features. Lee and Myaeng (2002) use word statistics to determine which terms are most discriminative between genres and between

topics. Finally, Finn and Kushmerick (2006) compare the bag-of-words approach with simple text statistics and conclude that both methods achieve high classification accuracy on fixed topic-genre combinations but perform worse when predicting topic-independent genre labels.

While most genre classification research has focused on the English language, some work has addressed language-independent, i.e., mono-lingual methods applicable to any language (Sharoff, 2007; Sharoff et al., 2010), or cross-lingual genre classification, i.e., training a classifier on one language to use it for classification in another language (Gliozzo and Strapparava, 2006; Petrenz, 2012; Petrenz and Webber, 2012), indicating that a single type of features can often generalize across multiple languages. In this chapter we develop language-independent genre classifiers. While the exact features differ among languages, we provide a single recommendation for genre classification regardless of the language.

4.2.3 Adaptation using classifiers

Adaptation using automatic classifiers has been applied in a few previous efforts. For example, Xu et al. (2007) build a general PBMT system with domain-specific tuning and language modeling. To decide which domain-specific system has to be used for an incoming document, they use text classification techniques based on LM perplexity or bag-of-words similarity. Banerjee et al. (2010) apply a similar method, using an SVM classifier with *tf-idf* word bigram features. Translation models in their work are built using in-domain data and tuning. Finally, Wang et al. (2012) also combine text classification with routing test documents to the most appropriate decoder settings. In addition, they find that BLEU loss strongly correlates with classification errors.

All three methods limit their application to two provenance-based domains, which are typically very distinct, e.g., patents versus 'generic'. To the best of our knowledge, we are the first to extend this setup up to four genres and four language pairs, where documents within a genre originate from a variety of sources.

4.3 Bilingual resource acquisition

We want to study the impact of genre differences on PBMT on a diverse set of language pairs, covering both commonly and rarely studied language pairs. Parallel training data for common language pairs is abundant but limited to a few genres such as parliamentary and legal proceedings. For low-resource languages the situation is—by definition—much worse, with very few to no bilingual corpora available. In addition, we found in Chapter 3 that poor translation quality for genre differences can to a large extent be attributed to poor model coverage. We therefore start our work in this chapter by describing how we automatically harvest parallel data from the web.

In this section we describe the language pairs and genres of interest (§4.3.1), the technical details of the data automated collection process including the main web sources that we used for data collection (§4.3.2), and the specifications of the final training and evaluation sets (§4.3.3).

4.3.1 Language pairs and genres

While most research in MT is evaluated on a small number of high-resource language pairs and 'domains', we opt for a more balanced distribution of source languages that allows us to measure to what extent our findings for common language pairs generalize to languages with limited resources. We therefore evaluate our experiments in this chapter on the following language pairs:

Arabic-English (AR-EN) is commonly used in research literature and often covered in MT evaluation campaigns. Arabic is a morphologically rich language, and has local to medium-long word-reordering with English. Sizable Arabic-English resources exist covering mostly news and UN documents.

Chinese-English (ZH-EN) is commonly used in research literature and often covered in MT evaluation campaigns. Chinese has no morphology, but complex word-reordering with respect to English. Sizable Chinese-English resources exist covering mostly news and parliamentary proceedings.

Bulgarian-English (BG-EN) is not often used in research literature and not covered in MT evaluation campaigns. Bulgarian has some morphology and only short-distance word-reordering with respect to English. Bulgarian-English resources are limited to Europarl (Koehn, 2005).

Persian (Farsi)-English (FA-EN) is rarely used in research literature and was only once covered in an MT evaluation campaign. Persian is morphologically rich and has long-distance word-reordering with respect to English. There are no publicly available Persian-English resources.

For each of these language pairs we consider four different genres:

News covers general news, as it can be found in (online) newspapers, and in transcripts of broadcast news.

Editorial covers Op-Ed pieces in (online) newspapers, that represent a subjective, and unlike news less matter-of-fact point of view. Editorials often consist of complex arguments, assuming a certain level of sophistication from their readership.

Colloquial covers informal conversation such as blog comments and Internet forum discussions. This is the most conversational and interactive genre in this list.

Speech covers speeches from which transcripts are available. Note that this category does not include informal spoken language, and does not require speech recognition tools.

While not all language-genre combinations are equally common, we can construct at least a translation test set for each of the four genres in each of the four language pairs. To this end, we use a large number of web sources, see Table 4.1.

Table 4.1: Overview of web sources used to collect bilingual corpora for AR-EN, BG-EN, FA-EN, and ZH-EN.

	Contains documents in				
Web source	AR-EN	BG-EN	FA-EN	ZH-EN	Available genre(s)
globalvoicesonline.org	✓	✓	✓	✓	news, colloquial
mawtani.al-shorfa.com	✓		✓		news, colloquial
al-shorfa.com	✓		✓		news, colloquial
magharebia.com	✓				news, colloquial
sabahionline.com	✓				news, colloquial
www.setimes.com		✓			news, colloquial
centralasiaonline.com			✓		news, colloquial
www.sada-e-azadi.net			✓		news
www.tolonews.com			✓		news
iwpr.net	✓		✓		news
www.niqash.org	✓				news
www.fao.org	✓			✓	news
www.cuyoo.com				✓	news
cn.nytimes.com				✓	news, editorial
www.worldbank.org	✓	✓	✓	✓	news, editorial, speech
translations.state.gov	✓		✓	✓	news, editorial, speech
www.chinadialogue.net				✓	editorial, colloquial
www.danielpipes.org	✓	✓	✓	✓	editorial
www.project-syndicate.org	✓			✓	editorial
www.ted.com	✓	✓	✓	✓	speech

4.3.2 Systematic data harvesting

Crawling and extracting textual content from the web is a time-consuming procedure as it has to be adapted to different formats for different sources and changes of format over time. To make this procedure more flexible and less time-consuming we create a single software component that employs a structured and easy-to-use data harvesting method that works for all of the above listed web pages listed in Table 4.1. The software is written in Python and makes use of *BeautifulSoup*[2], a library that extracts data from HTML files. The only adjustment needed for different web pages is a source-specific tag file as input to the data harvesting software. See Figure 4.1 for an example.

The software component follows a simple work flow. First, the source-specific tag

[2]http://www.crummy.com/software/BeautifulSoup/

```
@info: SOURCE           - sabahionline
@info: GENRE            - news
@info: HOMEPAGE         - http://sabahionline.com
@info: OVERVIEW         - /en_GB/archives/2013/07
@info: DATE_FORMAT      - year month day
@head: HEADLINE         - meta[name=title]
@head: DESCRIPTION      - meta[name=description]
@head: DATE             - meta[name=date]
@body: BODY             - div[id=articlescontent]
@body: CAPTIONS         - li[class=image_block]
@user: COMMENTS         - div[id=comments_comments]
```

Figure 4.1: Example of source-specific tags. The first field indicates how a tag will be processed in the software: `@info` tags contain meta information of the web source. `@head` and `@body` tags are css selectors used for extracting information form the head and body of the HTML page, respectively. `@user` tags indicate the location of user-generated data in an HTML file. All css selector values can be replaced by 'none' if the specific tag is not present on a web page.

file is read and processed. This file contains the URLs of one or more overview pages, such as news archives or sitemaps, which are then downloaded. Next, a list of URLs of actual text documents is extracted from each overview page. After downloading these pages, we then identify their translations in one of the following two ways: we first try to follow direct links, such as 'read this article in Arabic.' If no such links exist, we replace the language abbreviation in the url. For example, `cn.nytimes.com/20130922/kenya/zh-hant` contains a Chinese translation of `cn.nytimes.com/20130922/kenya/en-us`. Following this procedure, documents with no direct links will not be marked as translations. However, with only a few exceptions, all parallel pages are easily detected using these criteria.

4.3.3 Training and evaluation sets

Our data harvesting efforts yield large numbers of harvested bilingual documents for many genres and language pairs. However, PBMT system training requires data to be sentence-parallel rather than document-parallel. Hence, we have to align sentences that are translations of each other from parallel documents. To this end, we use a combination of two sentence alignment techniques:

- **Moore's sentence alignment:** An alignment method based on a combination of sentence length and word correspondence (Moore, 2002). This alignment approach does not require any prior knowledge of the languages in the corpus, but comes with a computational cost.
- **Champollion sentence alignment:** A sentence alignment approach that requires a dictionary of the target and source languages in a language pair, but is computationally easier and more robust than Moore's sentence alignment (Ma, 2006).

In this work we first use Moore's sentence aligner on a subset of the parallel documents. During this procedure a dictionary is created, which we then use as input for the Champollion sentence aligner. This combinatorial approach optimizes both sentence alignment quality and computational cost.

Table 4.2: Specifications of the harvested Arabic-English training, development and test data. Tokens are counted on the English side of the corpora.

	Training data			Development set			Test set		
Genre	Docs	Lines	Tokens	Docs	Lines	Tokens	Docs	Lines	Tokens
Colloquial	18.6K	273K	8.9M	89	1.5K	77.3K	110	1.5K	73.0K
Editorial	5.1K	156K	4.7M	46	1.5K	45.6K	47	1.5K	47.3K
News	59.7K	600K	18.0M	175	1.5K	50.4K	168	1.5K	48.1K
Speech	1.7K	140K	3.4M	20	1.5K	35.7K	19	1.5K	38.7K
Total	85.1K	1.2M	35.0M	330	6.0K	209K	344	6.0K	207K

Table 4.3: Specifications of the harvested Chinese-English training, development and test data. Tokens are counted on the English side of the corpora.

	Training data			Development set			Test set		
Genre	Docs	Lines	Tokens	Docs	Lines	Tokens	Docs	Lines	Tokens
Colloquial	4.0K	55K	1.7M	84	1.5K	42.5K	104	1.4K	35.8K
Editorial	10.1K	370K	10.2M	42	1.5K	43.1K	44	1.5K	42.6K
News	20.5K	584K	16.4M	68	1.5K	39.2K	50	1.5K	35.8K
Speech	1.7K	146K	3.3M	33	1.5K	42.6K	33	1.5K	37.5K
Total	36.3K	1.2M	31.6M	227	6.0K	169K	231	5.9K	152K

Next, we organize the collected bilingual data into training, development, and test sets. To avoid dependencies between the different data portions, we compose them by selecting documents from fixed time intervals, a strategy that is also used by the LDC. In our case, all training documents are older then the evaluation documents, and the development documents are older than the test documents. Obviously, there is no temporal overlap between any of the data portions. For evaluation, we select documents from a one-year period. The first six months are used for generating the development set and the second six months are used for generating the test set.

We tokenize all Arabic data using MADA, segment the Chinese data following Tseng et al. (2005), and use a simple in-house tokenizer for the other languages. The total numbers of foreign-English sentence pairs for our four genres in four language pairs are listed in Tables 4.2–4.5.

4.4 Genre-specific baseline systems

We use the newly assembled resources to create PBMT systems tailored to translate four different genres. In this section we first discuss the general experimental setup (§4.4.1) and then train and evaluate baseline PBMT systems for each individual genre and for a mixture of genres (§4.4.2).

Table 4.4: Specifications of the harvested Bulgarian-English training, development and test data. Tokens are counted on the English side of the corpora.

	Training data			Development set			Test set		
Genre	Docs	Lines	Tokens	Docs	Lines	Tokens	Docs	Lines	Tokens
Colloquial	–	–	–	–	–	–	146	1.4K	33.9K
Editorial	–	–	–	–	–	–	6	178	5.1K
News	8.4K	215K	5.3M	57	1.2K	30.2K	68	2.0K	49.5K
Speech	1.5K	206K	3.9M	16	1.2K	22.5K	22	2.0K	44.6K
Total	9.9K	422K	9.2M	73	2.4K	52.7K	242	5.6K	133K

Table 4.5: Specifications of the harvested Persian-English training, development and test data. Tokens are counted on the English side of the corpora.

	Training data			Development set			Test set		
Genre	Docs	Lines	Tokens	Docs	Lines	Tokens	Docs	Lines	Tokens
Colloquial	15.9K	629K	16.4M	19	1.5K	40.3K	25	1.5K	37.7K
Editorial	–	–	–	–	–	–	19	600	19.4K
News	49.1K	618K	16.8M	148	1.5K	44.5K	163	1.5K	47.4K
Speech	1.2K	119K	2.5M	14	1.5K	31.2K	18	1.5K	35.6K
Total	66.2K	1.4M	35.7M	181	4.5K	116K	225	5.1K	140K

4.4.1 Experimental setup

All systems in this chapter are trained using Oister (Section 2.4.1). We use the data harvested from the web (Section 4.3), supplemented with commonly used training data (if available) such as LDC corpora for Arabic-English and Chinese-English, and Europarl data (Koehn, 2005) for Bulgarian-English. The Persian-English system uses exactly the same training data as listed in Table 4.5. Training data specifications of the Arabic-English, Chinese-English, and Bulgarian-English systems are shown in Tables 4.6, 4.7 and 4.8, respectively.

In order to create genre-specific PBMT systems, we have to adequately use the available data. Simply concatenating the different corpora yields a general PBMT system that performs reasonably well across a variety of genres, i.e., those covered in the training data, but is not optimal for each individual genre. Since we aim to create genre-specific systems, we use a fill-up technique (Bisazza et al., 2011), in which we combine models trained on a particular genre with models trained on the remaining training corpora. Using this model combination technique, an additional feature is learned that favors genre-specific models, and 'backs off' to additional (out-of-genre) models for phrases that are unseen in the genre of interest. For instance, to train our news translation system, we train two phrase tables: one using all news data and one using all non-news data. We use the latter to complement the first with phrase pairs that are not covered in the first.

Table 4.6: Specifications of the complete Arabic-English training data. Detailed statistics of the web crawls are listed in Table 4.2.

Corpus/genre	Sentence pairs	English tokens
LDC broadcast conversation	47K	1.3M
LDC broadcast news	41K	1.1M
LDC (comparable) newswire	816K	26M
LDC weblogs & newsgroups	23K	577K
LDC SMS/chat/CTS	136K	1.6M
Web crawls (Table 4.2)	1.2M	35M
Total	2.2M	66M

Table 4.7: Specifications of the complete Chinese-English training data. Detailed statistics of the web crawls are listed in Table 4.3.

Corpus/genre	Sentence pairs	English tokens
LDC broadcast conversation	67K	1.2M
LDC broadcast news	69K	1.8M
LDC (comparable) newswire	1.0M	32M
LDC weblogs & newsgroups	45K	1.2M
LDC SMS/chat/CTS	181K	2.2M
LDC lexicons	1.2M	2.0M
Web crawls (Table 4.3)	1.2M	32M
Total	3.7M	72M

Table 4.8: Specifications of the complete Bulgarian-English training data.

Corpus/genre	Sentence pairs	English tokens
Europarl	393K	11M
Web-crawled news	215K	5.3M
Web-crawled speech	206K	3.9M
Total	815K	20M

Following the above strategy we can train genre-specific systems for all genres for which we have training data. Genres not covered in the training data have to be translated using a system trained on a mixture of genres or on one of the other genre-specific systems. For example, editorial Persian-English data is scarce, so for Persian editorial documents we have to resort to our colloquial, news, speech or mix system.

In addition to using the fill-up approach, we tune each genre-specific system on a development set covering only the genre of interest.

Table 4.9: Arabic-English translation quality in BLEU of four test genres using different genre-optimized systems and a genre-agnostic baseline. Best results for each test set genre are boldfaced. 'Combined best BLEU' indicates the overall BLEU score when combining the bold-faced results of all test genres in a single test set, followed by the difference with the genre-agnostic system.

Test set genre	Genre-agnostic baseline	PBMT system optimized for				Combined best BLEU
		colloquial	editorial	news	speech	
Colloquial	11.7	**13.8**	10.8	11.7	11.2	**17.9** (+1.1)
Editorial	22.6	19.6	**23.5**	21.6	21.0	
News	22.6	20.2	21.7	**23.2**	21.2	
Speech	11.5	11.5	11.1	11.0	**11.7**	
All	16.8	16.6	16.4	16.6	16.0	

Table 4.10: Chinese-English BLEU of various systems, see Table 4.9 for detailed explanation.

Test set genre	Genre-agnostic baseline	PBMT system optimized for				Combined best BLEU
		colloquial	editorial	news	speech	
Colloquial	11.4	**11.6**	11.3	10.7	11.3	**13.9** (+0.5)
Editorial	15.5	14.9	**16.3**	14.6	14.3	
News	13.3	12.8	13.3	**13.5**	12.4	
Speech	12.8	12.5	12.5	12.1	**13.9**	
All	13.4	13.1	13.4	12.7	13.2	

4.4.2 Results

Tables 4.9-4.12 provide translation quality results for all language pairs. For each language pair, we measure BLEU for our four test genres with the available genre-specific systems as well as the genre-agnostic system. Note that some Arabic-English and Chinese-English BLEU scores might be lower than those reported in literature since our test data contains only one reference translation, whereas some commonly used publicly available test sets have four reference translations.

The results confirm our expectation that the various test set genres benefit from being translated using a genre-optimized system rather than using a general system: generally, the highest BLEU scores are located on the diagonal of each table. In cases were no genre-specific system is available, we see that the best results are mostly obtained using the general system rather than a system optimized for a different genre.

Genre differences across language pairs. Next, in order to study patterns across language pairs and genres, we measure the linear correlation between the amount of genre-specific parallel data in the genre-agnostic system and the observed BLEU differences between genres, see Figure 4.2. If we do this for all language pairs at once (left in Figure 4.2), we find a Pearson correlation coefficient of $r = -0.04$. This indicates that over the four language pairs, there is no correlation between the amount

Table 4.11: Bulgarian-English BLEU of various systems, see Table 4.9 for detailed explanation.

Test set genre	Genre-agnostic baseline	PBMT system optimized for				Combined best BLEU
		colloquial	editorial	news	speech	
Colloquial	**29.1**	–	–	28.1	28.1	**33.4** (+0.6)
Editorial	24.7	–	–	**25.4**	21.3	
News	39.8	–	–	**40.4**	34.7	
Speech	27.4	–	–	25.8	**28.4**	
All	32.8	–	–	31.9	30.5	

Table 4.12: Persian-English BLEU of various systems, see Table 4.9 for detailed explanation.

Test set genre	Genre-agnostic baseline	PBMT system optimized for				Combined best BLEU
		colloquial	editorial	news	speech	
Colloquial	22.4	**22.5**	–	20.9	21.5	**22.3** (+0.4)
Editorial	**15.7**	15.2	–	15.6	15.1	
News	24.2	22.3	–	**24.3**	23.0	
Speech	21.3	19.5	–	20.7	**22.6**	
All	21.9	20.8	–	21.3	21.5	

of genre-specific parallel training data and BLEU. Instead, we see that BLEU scores vary more between language pairs than between genres, reflecting that some languages are easier to translate despite limited amounts of training data (e.g., Bulgarian-English) while others achieve poor BLEU scores even though large amounts of training data are available (e.g., Chinese-English).

However, if we look at training data and BLEU scores *within* each of the language pairs, we observe moderate to strong positive correlations, see Table 4.13. This is also visible in the right plot in Figure 4.2, where we apply per-language pair normalization of training data sizes and BLEU scores. Concretely, we measure and display for each genre (i) the proportion of genre-specific data in the training bitext, and (ii) the BLEU difference with respect to the average BLEU for the language pair. Using these normalized values we observe a Pearson correlation of $r = 0.61$, suggesting that differences in PBMT performance among genres can to some extent be explained by the proportion of genre-specific resources.

The right plot in Figure 4.2 also shows some patterns that are consistent among language pairs. For instance, both BLEU scores and resource proportions are always high for news, while translation quality for colloquial and speech documents is always relatively poor. Translation quality for editorial pieces varies: in the two language pairs (Arabic-English and Chinese-English) for which the editorial genre is covered in the parallel training data—even in small amounts—, BLEU scores are relatively high. However, when no editorial training data is available, BLEU scores are lowest of all four genres.

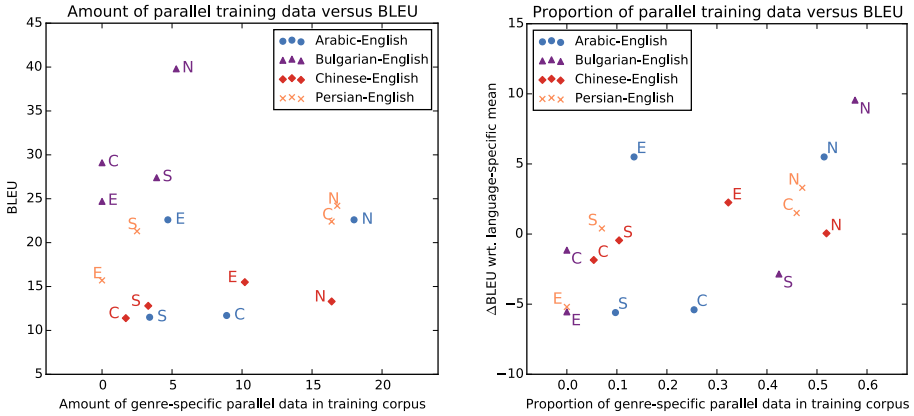


Figure 4.2: Training data size versus BLEU for four genres in four language pairs. Left: absolute bitext size and absolute BLEU show no linear correlation ($r = -0.04$). Right: relative (per language pair) bitext size and BLEU show a moderate positive correlation ($r = 0.61$). Genre abbreviations: C=colloquial, E=editorial, N=news, S=speech.

Table 4.13: Pearson correlation coefficients for four language pairs between BLEU scores and the amount of genre-specific parallel training data.

Language pair	Pearson's r
Arabic-English	0.46
Bulgarian-English	0.73
Chinese-English	0.55
Persian-English	0.83

4.5 Automatic genre classification

We observed that translation quality is usually best when translating each genre using its respective genre-specific baseline system. This motivates the hypothesis that translation of a mixture-of-genre test set can be improved by using a genre classifier, which routes test sentences or documents to the most appropriate PBMT system. Figure 4.3 illustrates this setup, which we put into practice in the current and the next section.

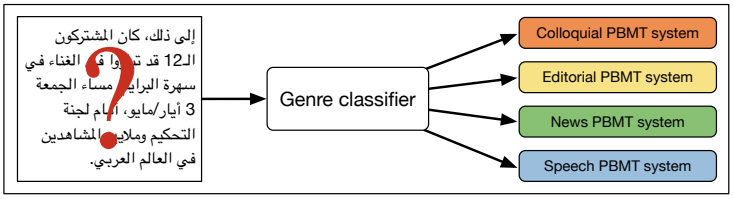


Figure 4.3: Schematic overview of our genre adaptation pipeline using genre classification. The genre classifier escorts an incoming document to the most relevant genre-specific PBMT system.

In this section we address the first component of this genre adaptation pipeline: building an accurate genre classifier. In particular, we aim at developing a single classification procedure that can be used on any source document regardless of the language it is written in. For this purpose, we apply our experiments to three languages: Arabic, Chinese, and English.

4.5.1 Building classifiers

We compare a number of classification techniques and feature sets, which we train and evaluate using a subset of the harvested data sets introduced in Section 4.3. Specifications of the data, features and classifiers are outlined below.

Data

For our genre classifiers we consider the same four text genres as covered in our data harvesting efforts: colloquial, editorial, news, and speech. To train the classifiers we randomly select documents from the training data listed in Tables 4.2 and 4.3. The complete selection comprises 1,000 documents per genre, thus enforcing equal prior classification probabilities for all genres. To obtain stable results, we repeat this selection strategy ten times, yielding ten different sets of 4×1000 training documents. All results reported in Section 4.5.2 are averaged over these ten training subsets.

We evaluate the performance of our classifiers in two ways: first, we apply 10-fold stratified cross-validation on the training data. Second, we report classification accuracy on held-out test documents, which are exactly the test documents used for Arabic-English and Chinese-English translation, see Tables 4.2 and 4.3. A summary of the test document statistics is given in Table 4.14.

Table 4.14: Test document statistics per language used for genre classification. The Arabic and Chinese test collections comprise the source sides of the parallel test documents listed in Tables 4.2 and 4.3. The English test collection comprises the target sides of these documents.

	Number of test documents			
Language	Colloquial	Editorial	News	Speech
Arabic	110	47	168	19
Chinese	104	44	50	33
English	214	91	218	52

Features

We experiment with three different features settings: first, we extract document features with the bag-of-words (BOW) approach, which is often used in subject or topic classification (Sebastiani, 2002; Moschitti and Basili, 2004). For this task, it is common to remove stopwords since these are not very distinctive for the subject of a document. However, for genre classification, we expect that stopwords do have the potential to distinguish between various genres, so we include these in our feature sets. Concretely,

we collect the set of n most common words per genre, and take the union of these genre-specific sets as our features. We repeat this setting for various values of n: $n \in \{10, 50, 100, 500, 1000, 5000\}$. We use relative frequencies rather than binary labels of the n most common words per document as our feature values.

Our second setting closely resembles the previous approach. However, instead of the words' surface forms we use part-of-speech (POS) tags as document features, which we extract using the Stanford POS tagger (Toutanova et al., 2003). In previous work, POS tags have been shown to be informative in text genre classification of English text (Finn and Kushmerick, 2006). Since there exist only a few dozens of different POS tags,[3] we do not select the most common ones, but we include them all. Feature values are again represented as relative frequencies.

In our third setting, we combine the previous two approaches. Here, the set of features includes the top n most common words as well as all POS tags. Previous work on text genre classification has shown that such combined feature sets result in high classification accuracy (Dewdney et al., 2001).

Models

For the classification experiments in this work, we use three types of classifiers: Naive Bayes (John and Langley, 1995), C4.5 Decision tree (Quinlan, 1993), and Support Vector Machines (SVM) using the sequential minimal optimization (SMO) algorithm (Platt, 1998) with linear kernels. SVMs in particular have shown to be very suitable for text classification tasks (Dewdney et al., 2001), often superior to other classifiers (Joachims, 1998). We run all experiments using the WEKA data mining software (Hall et al., 2009).

4.5.2 Results

Figures 4.4, 4.5, and 4.6 show the results of our classification experiments for Arabic, Chinese, and English, respectively. The line plots in the left columns show training, 10-fold cross-validation and test set accuracy for the three different classifiers using various BOW feature sets. These figures do not display POS or combined features, but exclusively compare various sizes of n in the BOW feature approaches, thus giving an indication of the optimal number of words. For all languages and classifiers, we observe that the most accurate predictions result from using 500 or 1000 word features. Increasing the number of words to 5000 either stabilizes prediction accuracy or—particularly when using SVMs—leads to overfitting of the model parameters.

The bar plots in the right columns of Figures 4.4–4.6 show a comparison on the test documents between using BOW features alone (green bars) or in combination with POS features (purple bars). In addition, the dashed lines indicate classification accuracy when using only POS features. In each figure, the best performing variant per classifier is marked with its accuracy. While results vary between language pairs and classifiers, we can formulate some general findings:

[3] The exact number depends on the language. On our training documents, the Stanford POS tagger finds 30 distinct tags for Arabic, 35 for Chinese, and 46 for English.

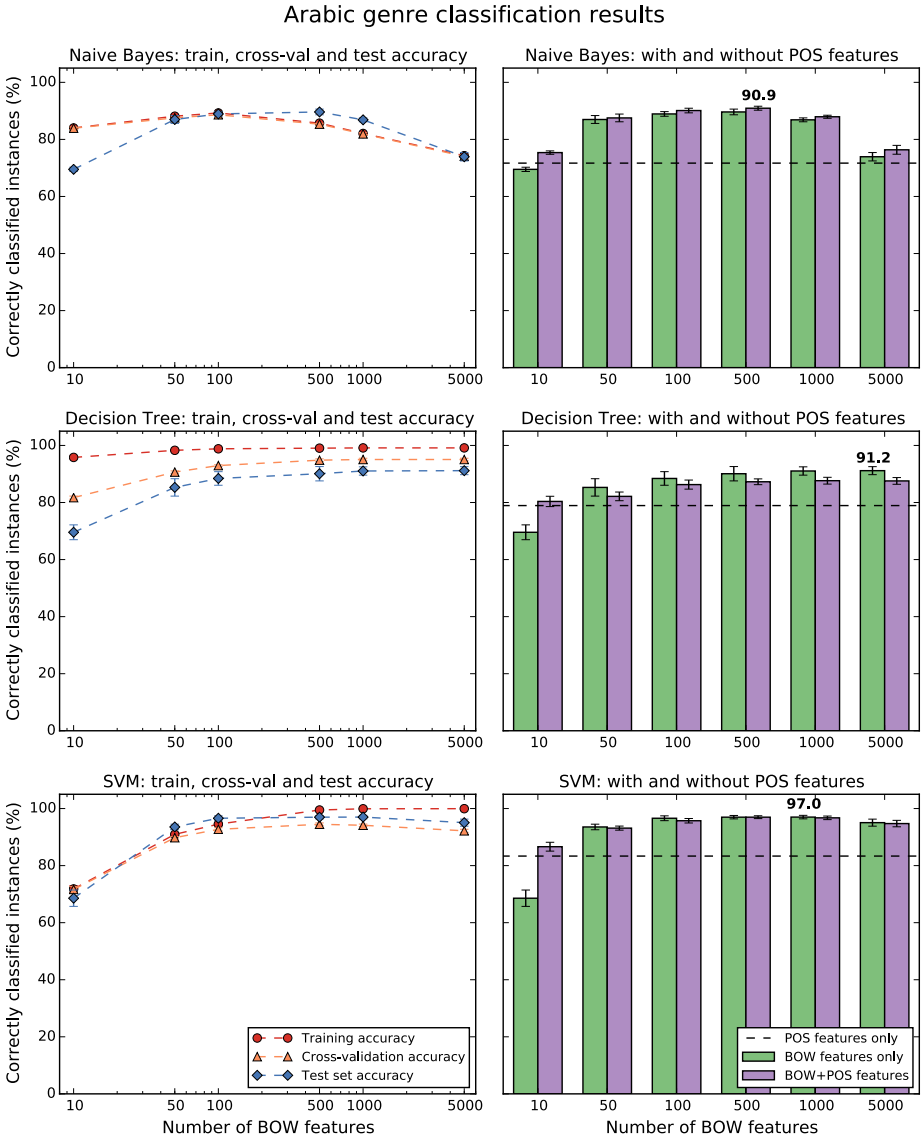


Figure 4.4: Arabic genre classification results (average of 10 classifiers) using Naive Bayes (top), C4.5 decision tree (center), and SVM (bottom) classifiers. Left: training, 10-fold cross-validation and test set accuracies of different BOW features sets. Right: Test set accuracies using various BOW feature sets excluding and including POS features. Dashed black line indicates test set accuracy when using only POS features. Value of the best performing feature set for each classifier is displayed above the corresponding bar. All error bars represent standard deviations over 10 training samples.

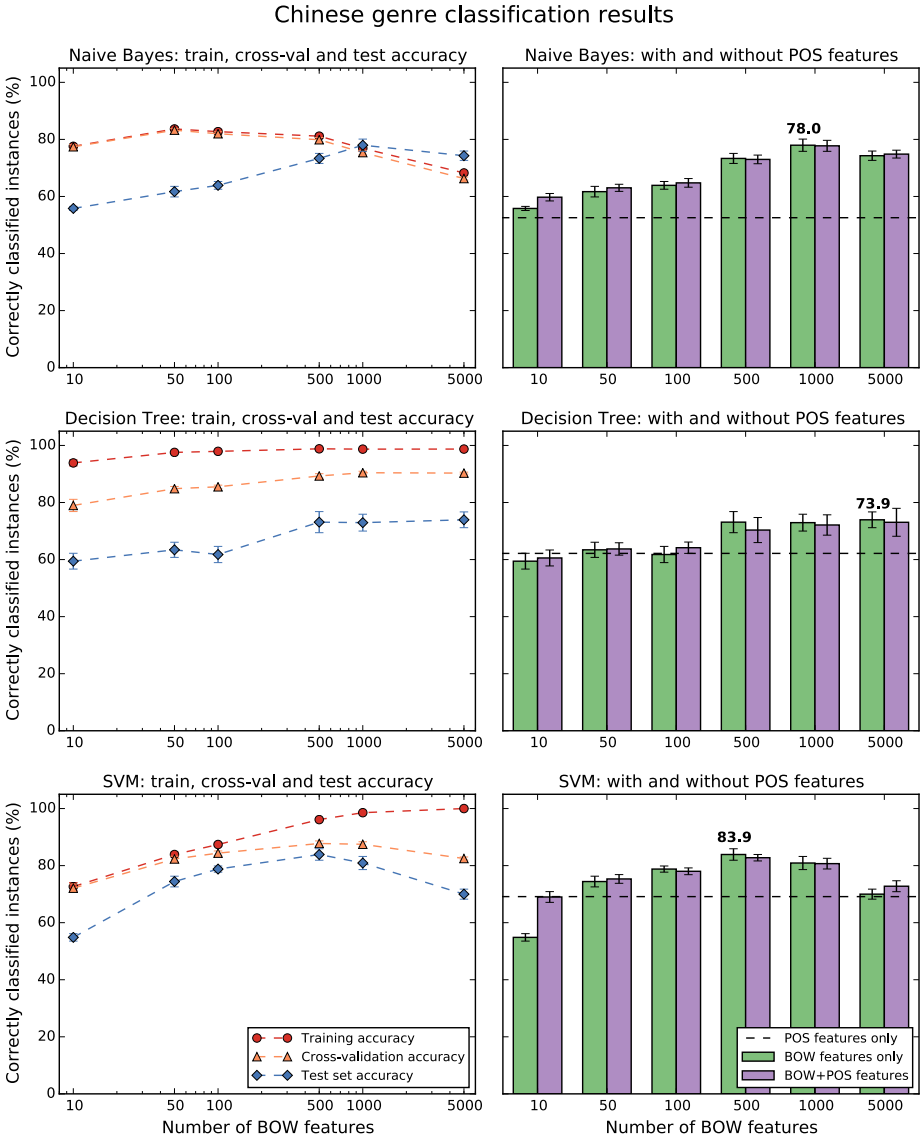


Figure 4.5: Chinese genre classification results (average of 10 classifiers) using Naive Bayes (top), C4.5 decision tree (center), and SVM (bottom) classifiers. Left: training, 10-fold cross-validation and test set accuracies of different BOW features sets. Right: Test set accuracies using various BOW feature sets excluding and including POS features. Dashed black line indicates test set accuracy when using only POS features. Value of the best performing feature set for each classifier is displayed above the corresponding bar. All error bars represent standard deviations over 10 training samples.

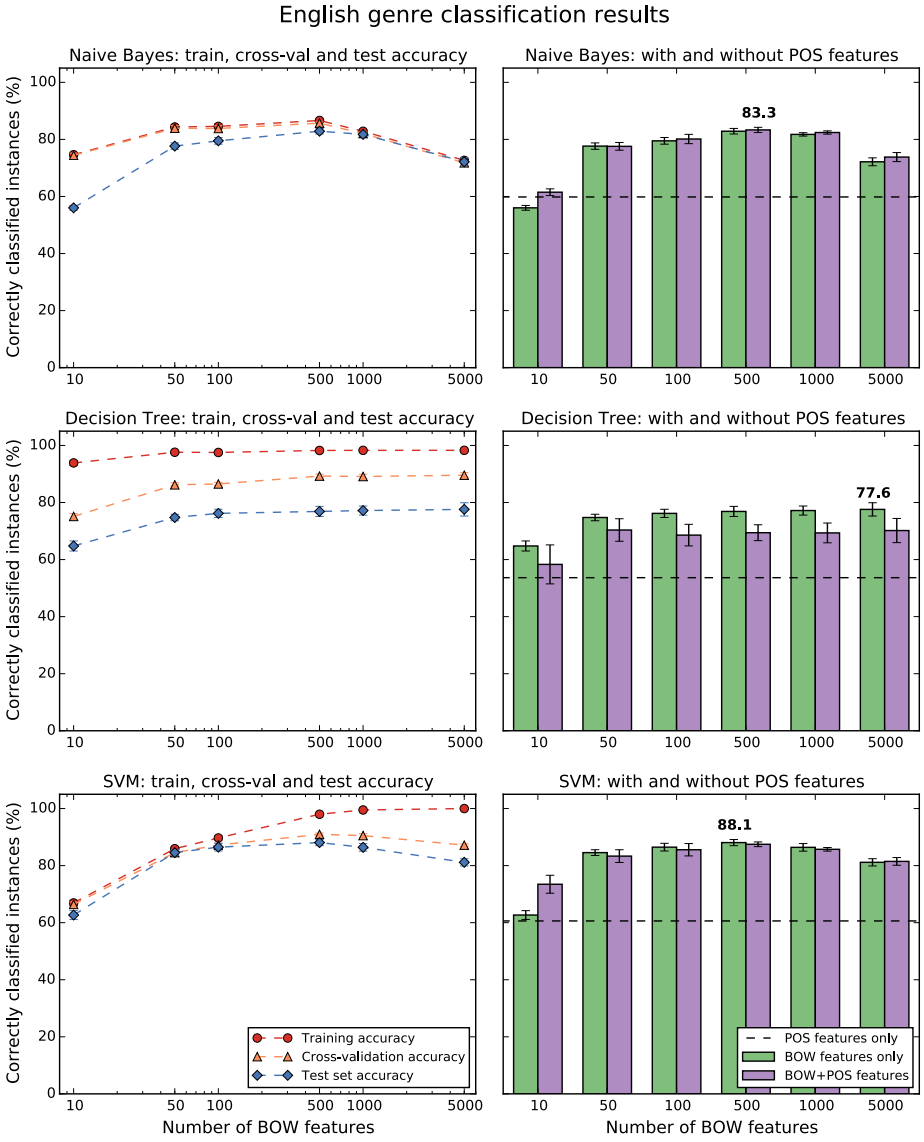


Figure 4.6: English genre classification results (average of 10 classifiers) using Naive Bayes (top), C4.5 decision tree (center), and SVM (bottom) classifiers. Left: training, 10-fold cross-validation and test set accuracies of different BOW features sets. Right: Test set accuracies using various BOW feature sets excluding and including POS features. Dashed black line indicates test set accuracy when using only POS features. Value of the best performing feature set for each classifier is displayed above the corresponding bar. All error bars represent standard deviations over 10 training samples.

- When using the Naive Bayes classifier, it is best to use the combined BOW+POS feature set, while for the other classifiers performance is superior if only BOW features are used.
- 'More is better' when using the C4.5 Decision tree classifier. Test set accuracies are highest using the largest tested number of BOW features (5000).
- The overall best performance is achieved using the SVM classifier with 500 BOW features, which is therefore the recommended setting for our translation experiments in the remainder of this chapter. An additional advantage of using BOW features only is that we do not depend on the availability of POS tagging software, which only exists for a limited number of languages.

Confusion matrices. The above discussed figures only provide overall classification accuracies. However, since we are dealing with a multi-class classification task, it is interesting to examine which genres are mostly classified correctly and which misclassifications are common. To this end, we show in Table 4.15 for each language a confusion matrix of the test documents using the SVM classifier with 500 BOW features. Reported numbers are averaged over classifiers trained on ten different training samples.

Table 4.15: Confusion matrices for genre classification (average of 10 classifiers) using the best performing genre classifier (SVM in combination with 500 BOW features), measured on Arabic (left), Chinese (center) and English (right) test documents. Genre mapping: **C**=colloquial, **E**=editorial, **N**=news, **S**=speech.

Arabic test docs		Classified as C	E	N	S	Total
True label	**C**	**102**	2	3	3	110
	E	0	**47**	0	0	47
	N	0	0	**167**	1	168
	S	0	0	1	**18**	19
Total		102	49	171	22	344

Chinese test docs		Classified as C	E	N	S	Total
True label	**C**	**74**	13	9	8	104
	E	0	**44**	0	0	44
	N	3	3	**44**	0	50
	S	1	0	0	**32**	33
Total		78	60	53	40	231

English test docs		Classified as C	E	N	S	Total
True label	**C**	**171**	15	14	14	214
	E	0	**88**	1	2	91
	N	7	11	**198**	2	218
	S	1	0	1	**50**	52
Total		179	114	214	68	575

For all languages, classification of editorial and speech documents is extremely accurate, sometimes even perfect. A likely explanation is that editorial and speech documents are on average longer than news and colloquial documents, resulting in more non-zero feature values which allow for more accurate class prediction. Colloquial documents, on the other hand, have the largest numbers of misclassifications, which can be explained by the fact that these documents are highly variable in their nature since they are not editorially controlled or standardized.

Most discriminative features. Finally, it is interesting to look at which features are most discriminative for each genre. Since our multi-class SVM classifier comprises multiple binary SVM classifiers, we show in Table 4.16 the top five most important English BOW features for both genres in each possible genre pair. From this table we can make a few interesting observations:

Table 4.16: Top five most important features for pairwise (each genre versus each other genre) English SVM genre classifiers that together make up the multi-class genre classifier. All features are tokenized and lowercased BOW features.

Colloquial vs. Editorial		Colloquial vs. News		Colloquial vs. Speech	
i	commentary	i	said	. . .	– –
this	distinguished	all	added	is	bank
think	analysis	this	post	should	laughter
is	recent	you	photo	any	percent
they	–	#	recent	am	applause

Editorial vs. News		Editorial vs. Speech		News vs. Speech	
,	said	commentary	– –	said	– –
but	blog	highest	percent	magharebia	we
commentary	bloggers	indeed	we	says	our
that	says	but	bank	blog	this
indeed	added	analysis	you	according	these

- Some words are very indicative for a genre regardless of the contrastive genre. For example, '– –' is always the most discriminative token for speech documents, 'said' is always the most discriminative token for news documents, and 'commentary' is always in the top three of most discriminative words for editorial documents.

- Some words are indicative for a genre only when it is contrasted with a particular other genre. For example, 'i' is the most discriminative word for colloquial documents except for classification of colloquial and speech since the latter also contains relatively high frequencies of 'i'.

- Some words can be indicative of multiple genres, also depending on which other genre is classified. For example, 'recent' is indicative for both editorial and news documents when contrasted with colloquial documents, however not in other scenarios. This suggests that the real information comes from the absence of 'recent' in colloquial documents rather than its presence in news or editorial documents.

- Overall, most of the top five BOW features are stopwords, justifying our choice to not remove stopwords for genre classication.

To summarize, we have shown in this section that we can learn accurate genre classifiers for four text genres using SVM classifiers and a feature set containing the 500 most common words per genre. This classification setup works well for at least three languages; Arabic, Chinese, and English, and is thus very suitable for PBMT adaptation, where the input documents are often not written in English.

4.6 Genre adaptation using automatic classifiers

In this section we address the second component of the genre adaptation pipeline illustrated in Figure 4.3: incorporating genre classification into an end-to-end PBMT system. Concretely, we classify for each document in the test set its genre, and guide it to the most appropriate PBMT system. Note that while we do have access to the true genre labels in this controlled research scenario, we intentionally mimic a more realistic situation in which an incoming test sentence is of unknown origin.

Figures 4.17–4.20 show the translation quality measured with BLEU for all language pairs using (i) a genre-agnostic baseline system trained and tuned on a mixture of genres, (ii) several genre-specific systems which we combine manually and refer to as our 'oracle' system, and (iii) several genre-specific systems which we combine using automatic genre classifiers. For all subsequent experiments we use SVM classifiers with 500 BOW features since this proved the best performing setup in the previous section. We measure statistical significance with respect to the genre-agnostic baseline using approximate randomization (Riezler and Maxwell, 2005). We report significant improvements at the $p \leq 0.05$ ($\triangle$) or $p \leq 0.01$ ($\blacktriangle$) level, as well as significant deteriorations at the $p \leq 0.05$ ($\triangledown$) or $p \leq 0.01$ ($\blacktriangledown$) level.

Table 4.17: Arabic-English translation results in BLEU. Manual oracle results are combined from several genre-optimized systems using manual genre labels of the test documents, see Table 4.9. Improvements indicated with ▲ are statistically significant at the $p \leq 0.01$ level.

	System		
Genre	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	11.7	13.8$^{\blacktriangle}$	13.8$^{\blacktriangle}$
Editorial	22.6	23.5$^{\blacktriangle}$	23.5$^{\blacktriangle}$
News	22.6	23.2$^{\blacktriangle}$	23.2$^{\blacktriangle}$
Speech	11.5	11.7	11.6
Overall	16.8	17.9$^{\blacktriangle}$	17.8$^{\blacktriangle}$

Table 4.18: Chinese-English translation results in BLEU. Manual oracle results are combined from several genre-optimized systems using manual genre labels of the test documents, see Table 4.10. Improvements indicated with ▲ are statistically significant at the $p \leq 0.01$ level.

	System		
Genre	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	11.4	11.6	11.5
Editorial	15.5	16.3$^{\blacktriangle}$	16.3$^{\blacktriangle}$
News	13.3	13.5	13.6
Speech	12.8	13.9$^{\blacktriangle}$	14.0$^{\blacktriangle}$
Overall	13.4	13.9$^{\blacktriangle}$	13.9$^{\blacktriangle}$

Table 4.19: Bulgarian-English translation results in BLEU. Manual oracle results are combined from several genre-optimized systems using manual genre labels of the test documents, see Table 4.11. Statistically significant differences are indicated with $\triangle$ or $\triangledown$ at the $p \leq 0.05$ level and with ▲ or ▼ at the $p \leq 0.01$ level.

	System		
Genre	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	29.1	29.1	28.6$^{\blacktriangledown}$
Editorial	24.7	25.4$^{\triangle}$	25.4$^{\triangle}$
News	39.8	40.4$^{\blacktriangle}$	40.4$^{\blacktriangle}$
Speech	27.4	28.4$^{\blacktriangle}$	28.4$^{\blacktriangle}$
Overall	32.8	33.4$^{\blacktriangle}$	33.1$^{\blacktriangle}$

Table 4.20: Persian-English translation results in BLEU. Manual oracle results are combined from several genre-optimized systems using manual genre labels of the test documents, see Table 4.12. Improvements indicated with ▲ are statistically significant at the $p \leq 0.01$ level.

	System		
Genre	Genre-agnostic	Manual oracle	Genre-classified
Colloquial	22.4	22.5	22.5
Editorial	15.7	15.7	15.6
News	24.2	24.3	24.2
Speech	21.3	22.6$^{\blacktriangle}$	22.6$^{\blacktriangle}$
Overall	21.9	22.3$^{\blacktriangle}$	22.1$^{\blacktriangle}$

For Arabic-English and Chinese-English (Tables 4.17 and 4.18, respectively), we train our classifiers on four genres with a balanced prior distribution. Our Arabic genre classifier achieves near-perfect classification accuracy (97%), which is reflected by BLEU scores that are very similar—if not equal—to the oracle system. Our best Chinese genre classifier yields lower accuracy (84%), however BLEU scores of the genre-classified system do not suffer from this sub-optimal classification performance. On closer inspection we see that some documents actually benefit from being translated by a different genre-optimized system. For example, the three Chinese news documents classified as editorial (see confusion matrix in Table 4.15), totalling 189 sentences, improve with 0.4 BLEU if translated using the editorial rather than the news system. This may be explained by the nature of these documents: all three documents are longer than the average Chinese-English news article (which is 30 lines, see Table 4.3) and resemble editorial pieces in terms of writing style and vocabulary.

Next, we look at the languages for which not all genres are covered in the training data. Consequently, we can only train classifiers for two (Bulgarian) or three (Persian) of the genres in the test set. For the remaining test genres, the predicted genre will be one of the genres in the training data, and translation is performed using the corre-

sponding genre-specific system. Note that the genre-agnostic baseline system is never recommended based on classifier predictions, despite sometimes being the best option.

Table 4.19 shows the end-to-end results for Bulgarian-English translation. The Bulgarian genre classifier achieves 100% accuracy on news and speech, for which the classifier is trained. The six editorial test documents are all classified as news, which is advantageous for the PBMT output quality since the news system yields much higher BLEU than the speech system on the editorial translations (25.4 versus 21.3, see Table 4.11). Genre predictions for the 146 colloquial test documents are distributed fairly evenly over news and speech, achieving a BLEU score of 28.6. While the genre-agnostic system performs better (29.1), the result using an automatic classifier is superior to translating all colloquial documents with either the news (28.1) or the speech (28.0) system. This finding indicates that automatic genre classification can even be profitable if no training data for a given genre is available.

Finally, Table 4.20 shows the end-to-end results for Persian-English translation. The Persian classifier achieves 90% accuracy on the genres covered in the training data; colloquial, news, and speech. Of the 19 editorial documents, 14 are classified as news and 5 as speech. However, BLEU scores using the genre-agnostic and the genre-optimized systems are very similar for all genres except speech. Improvements using the genre-classified system are therefore small.

4.7 Conclusions

Motivated by the observation in Chapter 3 that genre differences pose a bigger challenge to PBMT than topic differences, we further explored in this chapter the impact of different genres and language pairs on PBMT quality.

Concretely, we extended our research in this chapter to four genres (colloquial, editorial, news, and speech) and four language pairs (Arabic-English, Bulgarian-English, Chinese-English, and Persian-English), for which we harvested parallel training, development and testing corpora from the web. The presented process of data harvesting is fully automated, and can easily be adapted to other web sources covering human translations.Armed with our new resources, we asked:

RQ2 *Is the observed impact of genre differences on PBMT consistent among various language pairs and data settings?*

To answer this questions, we first trained genre-agnostic baseline systems for each language pair. We then measured translation quality in BLEU as a function of the proportion of genre-specific resources in the training data (i) for all genres in all language pairs, and (ii) per language pair. While we found that variations in BLEU are to a large extent language-specific, we also observed moderate to strong positive correlations between genre-specific proportions of training data and BLEU scores. This indicates that differences in PBMT performance can to a certain degree be explained by the amount of available training data. Moreover, a number of genre-specific patterns generalize across language pairs and are thus no incidental observations. For example, news is always relatively easy to translate, whereas colloquial and speech documents achieve relatively low BLEU scores for all language pairs.

After studying the impact of genre differences on PBMT, we moved to genre adaptation for PBMT, asking:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

We exploited our new corpora to train genre-specific rather than genre-agnostic PBMT systems and tested the hypothesis that automatic genre classifiers can replace the need to manually guide test documents. To this end, we first trained genre classifiers for several languages. We achieved the best performance using an SVM classifier with the 500 most common words per genre as features. We then incorporated this genre classifier into an end-to-end PBMT pipeline, and showed that we can indeed for all four language pairs significantly improve translation quality over the genre-agnostic baseline system, without any dependency on knowing test set genre labels in advance. We also found that misclassification does not always lead to deterioration of translation quality, but rather benefits some documents. This indicates that manual genre labels do not always provide the most useful information for PBMT adaptation.

The classifiers used in this work are all crisp classifiers, i.e., they assign a single label per test document. Another strategy beyond the scope of the current chapter is to apply multi-label classification, in which different genre-specific PBMT models can be interpolated using classifier output weights. Such a setting might particularly benefit documents with low classification confidence, which are currently sent to a single system which may not be the most optimal.

5 Genre Adaptation Using Text Features

5.1 Introduction and research questions

In Chapter 3, we explicitly noted that *topic* and *genre* are two distinct properties of documents (Lee and Myaeng, 2002; Stein and Meyer Zu Eissen, 2006), and we believe that PBMT adaptation can be applied to both properties separately rather than only to the ambiguous concept *domain*. Specifically, we are interested in applying adaptation without relying on the availability of provenance information or manual subcorpus labels since these restrictions limit the applicability of adaptation methods. For example, we saw in Chapter 4 that misclassification of a document's genre can actually improve PBMT quality, indicating that manual genre labels do not always provide the most useful information for PBMT adaptation. Moreover, obtaining accurate subcorpus or genre labels may be time-consuming and not straightforward, as we will demonstrate later in this chapter.

While adaptation methods without subcorpus dependencies have been applied to topic adaptation, they have not been used yet to adapt PBMT systems to different genres, which is interesting since the latter pose larger challenges to PBMT than topics (see Chapter 3). We therefore ask:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

In Chapter 4 we addressed genre adaptation by means of model combination using genre automatic classifiers. In this chapter we explore genre adaptation with an instance weighting approach. Concretely, we replace its dependency on subcorpus labels with features that can be automatically extracted from the parallel bitext. First, we ask:

RQ3b. *Can we successfully adapt PBMT systems using textual indicators of genre?*

To this end, we incorporate genre-revealing features inspired by previous findings in the text genre classification literature, e.g., punctuation marks and different types of pronouns, into a competitive translation model adaptation approach with the aim of improving translation quality across two test sets with two genres: the first containing

newswire (NW) and user-generated (UG) comments, and the second containing NW and UG weblogs.

Next, we explore the use of latent Dirichlet allocation (LDA) features for the task of genre adaptation, and we ask:

RQ3c. *Can we successfully adapt PBMT systems using LDA features?*

LDA uses the words observed in a document collection to infer a number of latent threads, and has been used successfully for topic adaptation (see Section 3.3.2). Compared to topics, text genres may be less specified by the content words in a document. Nevertheless, our work on text genre classification in Section 4.5 showed that also genres can be distinguished by vocabulary differences. We therefore hypothesize that LDA may be useful for genre adaptation. In addition, we also investigate the combined use of textual genre features and LDA features.

Organization. This chapter is organized as follows: In Section 5.2 we discuss related work. Next, in Section 5.3 we describe the adaptation framework that we use as well as three different feature variants (subcorpus labels, genre-revealing text features, and LDA features). After a brief description of the experimental setup in Section 5.4, we discuss in Section 5.5 the PBMT results using different genre adaptation variants. Finally, we measure to what extent consistency of output translations is increased using our genre adaptation approach in Section 5.6, before summarizing this chapter in Section 5.7.

5.2 Related work

Some previous work related to this chapter has been discussed in Sections 3.2 (domain adaptation), 3.3.2 (topic and genre adaptation) and 4.2.2 (genre classification). Below we discuss two lines of relevant work in more detail: instance weighting methods for PBMT adaptation and LDA.

5.2.1 Instance weighting for PBMT adaptation

Instance weighting methods are commonly used for adaptation in PBMT. Such methods aim to prioritize bilingual training instances that are most relevant to the development and test data while not completely discarding less relevant data (which is done in data selection methods). One of the first and most well-known instance weighting approaches is presented by Matsoukas et al. (2009), who represent sentences in the training corpus by a vector capturing provenance information, then estimate a weight for each sentence by optimizing an objective function on the development set.

Most subsequent work learns relevance weights for phrase pairs rather than sentences. For example, Foster et al. (2010) use features that capture the usefulness of phrase pairs, e.g., perplexities, word frequencies and OOV rates, rather than manual subcorpus labels. However, their method still relies on known corpus boundaries of in-domain and out-of-domain data. Chen et al. (2013) propose a phrase pair weighting approach based on a vector space model (VSM), in which the vector dimensions are manually assigned subcorpus labels of the training data, and compute for each phrase

pair a similarity score with a vector representing the test domain. We discuss the technical details of this method in Section 5.3.1. Instead of computing a similarity score, Chen et al. (2014) add vector entries based on subcorpus frequencies as direct features to the phrase table, allowing for optimization per phrase pair but also substantially increasing the complexity and runtime during parameter tuning.

A number of dynamic variants of instance weighting have been proposed, which compute phrase pair relevance scores at decoding time rather than with respect to an in-domain development set (Costa-jussà and Banchs, 2011; Banchs and Costa-Jussà, 2011), and therefore alleviate the dependency on provenance information of the test set. Unfortunately, computing relevance scores of the entire training bitext for every test sentence under translation is a very time-consuming task. In this work, we want to explore to what extent different feature types can serve for genre adaptation and thus compare a large number of experiments. We therefore choose to work with a static rather than a dynamic instance weighting approach.

5.2.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA, Blei et al. (2003)) is a statistical model that infers hidden information that can explain similarity patterns in data. LDA is commonly used for topic modeling, in which it is assumed that each document describes a mixture of topical dimensions. These dimensions are, however, not known and are inferred probabilistically from the surface forms of the available documents. More formally, LDA assumes a generative probabilistic process from which the documents in a corpus arise. A graphical model illustrating this process is shown in Figure 5.1. Here each node represents a random variable, and each box represents a repeated process.

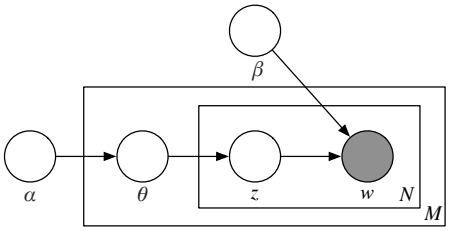


Figure 5.1: Graphical model representation of the generative process for document construction that is assumed in LDA. Figure adapted from Blei et al. (2003).

Concretely, the generative process consists of the following steps:

For each of M documents D_m in a corpus:

1. Given prior parameter α, choose a distribution θ over a mixture of topics.
2. For each of N word positions in document D_m:
    1. Given the topic distribution θ, assign a topic z.
    2. Given topic z and prior parameter β, choose a word w_n.

The documents in the corpus resulting from this generative process all share the same set of latent dimensions, but in different proportions, resulting in a different set of words for each document. Next, when facing a large document collection, we are not interested in the generative process, but in the inference of the hidden distributions θ for all documents. For this step, the algorithm needs to know the number of latent dimensions, and then searches for a latent distribution that best describes the observed documents.[1] Concretely, for each document, the LDA inference returns a distribution over the latent dimensions that probabilistically quantifies the relative contribution of each of the dimensions to the document during the assumed generative process.

Since LDA inference provides a way to automatically learn information about a document, it has been used to replace subcorpus or provenance information in topic adaptation (Tam et al., 2007; Eidelman et al., 2012; Hasler et al., 2012, 2014a,b,c; Hewavitharana et al., 2013; Su et al., 2015). However, we are interested in genre adaptation rather than topic adaptation. While LDA is a bag-of-words approach that may be insufficient to model genre accurately, we found in Chapter 4 that lexical features can in fact distinguish genres. We therefore explore in this chapter whether LDA can also be of use for genre adaptation.

5.3 Translation model genre adaptation

For the task of genre adaptation to the genres NW and UG comments or weblogs, we use a flexible translation model adaptation approach based on phrase pair weighting using a vector space model (VSM) inspired by Chen et al. (2013). The reason we choose an instance-weighting method rather than a mixture modeling approach is twofold: first, mixture modeling approaches intrinsically depend on subcorpus boundaries, which resemble provenance or require manual labeling. Second, Irvine et al. (2013a) have shown that including relevant training data in a mixture modeling approach solves many coverage errors, but also introduces substantial amounts of new scoring errors. With phrase-pair weighting we aim to optimize phrase translation selection while keeping our training data fixed, and we can thus compare the impact of several methodological variants on genre adaptation for PBMT.

5.3.1 VSM adaptation framework

In the selected adaptation method, each phrase pair in the training data is represented by a vector capturing information about the phrase:

$$V\left(\bar{f},\bar{e}\right) = \langle w_1\left(\bar{f},\bar{e}\right),\ldots,w_N\left(\bar{f},\bar{e}\right)\rangle. \tag{5.1}$$

Here, $w_i\left(\bar{f},\bar{e}\right)$ is the weight for phrase pair $\left(\bar{f},\bar{e}\right)$ of dimension $i \in N$ in the vector space. The exact definition of dimensions $i \in N$, and hence the information captured by the vector, depends on the definition of the vector space, for which we describe different variants in Sections 5.3.2–5.3.4.

[1] Typically, the inference methods are either *sampling-based* algorithms or *variational* algorithms. We will skip the technical details since their relevance for the purpose of adaptation in PBMT is very limited.

In addition to the phrase pair vectors, a single vector is created for the development set which is assumed to be similar to the test data:

$$V\left(dev\right) = \langle w_1\left(dev\right), \ldots, w_N\left(dev\right)\rangle, \tag{5.2}$$

where weights $w_i\left(dev\right)$ are computed for the entire development set, summing over the vectors of all phrase pairs that occur in the development set:

$$w_i\left(dev\right) = \sum_{\left(\bar{f},\bar{e}\right)\in P_{dev}} c_{dev}\left(\bar{f},\bar{e}\right) w_i\left(\bar{f},\bar{e}\right). \tag{5.3}$$

Here P_{dev} refers to the set of phrase pairs that can be extracted from the development set, $c_{dev}\left(\bar{f},\bar{e}\right)$ is the count of phrase pair $\left(\bar{f},\bar{e}\right)$ in the development set, and $w_i\left(\bar{f},\bar{e}\right)$ is the phrase pair's weight for dimension i in the vector space.

Next, for each phrase pair in the training corpus, we compute the Bhattacharyya Coefficient (BC, Bhattacharyya (1946)) as a similarity score[2] between its vector and the development vector:

$$BC\left(dev;\bar{f},\bar{e}\right) = \sum_{i=0}^{i=N} \sqrt{p_i\left(dev\right)\cdot p_i\left(\bar{f},\bar{e}\right)}, \tag{5.4}$$

where $p_i\left(dev\right)$ and $p_i\left(\bar{f},\bar{e}\right)$ are probabilities representing smoothed normalized vector weights $w_i\left(dev\right)$ and $w_i\left(\bar{f},\bar{e}\right)$, respectively.

The computed similarity is assumed to indicate the relevance of the phrase pair with respect to the development and test set and is added to the decoder as a new feature. In a similar fashion, two similarity-based decoder features $BC\left(dev;\bar{f},\bullet\right)$ and $BC\left(dev;\bullet,\bar{e}\right)$ are added for the marginal counts of the source and target phrases, respectively. Further technical details can be found in (Chen et al., 2013).

The presented framework for translation model adaptation allows us to empirically compare various sets of VSM features, of which we present three in the following sections.

5.3.2 Genre adaptation with subcorpus labels

First, we adhere to the commonly used scenario in which adaptation is guided by manual subcorpus labels that resemble provenance of training documents, which is also used by Chen et al. In this formulation, each weight $w_i\left(\bar{f},\bar{e}\right)$ in (5.1) is a *tf-idf* weight capturing the relative occurrence of phrase pair $\left(\bar{f},\bar{e}\right)$ in each subcorpus s_i in a total of C subcorpora:

$$w_i\left(\bar{f},\bar{e}\right) = tf_i\left(\bar{f},\bar{e}\right)\cdot idf\left(\bar{f},\bar{e}\right). \tag{5.5}$$

Here, *phrase pair frequency* $tf_i\left(\bar{f},\bar{e}\right)$ is computed as the raw count of phrase pair $\left(\bar{f},\bar{e}\right)$ in subcorpus s_i, divided by the maximum count of any phrase pair in s_i:

[2] Chen et al. compared three similarity measures and observed that the BC similarity performed best. We follow their recommendation in our work.

$$tf_i\left(\bar{f},\bar{e}\right) = \frac{c_i\left(\bar{f},\bar{e}\right)}{\max\{c_i\left(\bar{f}_j,\bar{e_k}\right) : \left(\bar{f}_j,\bar{e_k}\right) \in s_i\}}. \tag{5.6}$$

Next, the *inverse subcorpus frequency* $idf\left(\bar{f},\bar{e}\right)$ measures how common $\left(\bar{f},\bar{e}\right)$ is across all subcorpora:

$$idf\left(\bar{f},\bar{e}\right) = \log\left(\frac{C}{df\left(\bar{f},\bar{e}\right)} + \lambda\right). \tag{5.7}$$

Here, C is the number of subcorpora, $df\left(\bar{f},\bar{e}\right)$ refers to the *subcorpus frequency* and counts the number of subcorpora in which $\left(\bar{f},\bar{e}\right)$ occurs at least once, and λ is a smoothing term. Following Chen et al. we use $\lambda = 8$.

Since our aim is to adapt to multiple genres in a test corpus, we follow Chen et al. and manually group our training data into subcorpora that reflect various genres (see Table 5.3). While this definition of the vector space can approximate genres at different levels of granularity, manual subcorpus labels are labor intensive to generate, particularly in the scenario where provenance information is not available, and may not generalize well to new translation tasks.

5.3.3 Genre adaptation with genre features

To move away from manually assigned subcorpus labels, we explore the use of genre-revealing features that have proven successful for distinguishing genres in classification tasks (see Section 4.2.2). To this end, we construct a list of features that are directly observable in raw text, see Table 5.1. For each genre feature g_i, we first compute its raw count $c_i\left(d\right)$ per document d in the training corpus D, which we then normalize for document length:

$$c_i'\left(d\right) = \frac{c_i\left(d\right)}{|d|}, \tag{5.8}$$

where $|d|$ is the number of tokens in document d. We then scale each feature to a value in range $[0, 1]$ by min-max normalizing its values across all documents, and we obtain the final document-level feature value $w_i\left(d\right)$:

$$w_i\left(d\right) = \frac{c_i'\left(d\right) - \min\{c_i'(d_j) : d_j \in D\}}{\max\{c_i'(d_j) : d_j \in D\} - \min\{c_i'(d_j) : d_j \in D\}} \tag{5.9}$$

Next, each vector weight $w_i\left(\bar{f},\bar{e}\right)$ in (5.1) equals the weighted average of the document-level values of genre feature g_i for all training instances of phrase pair $\left(\bar{f},\bar{e}\right)$:

$$w_i\left(\bar{f},\bar{e}\right) = \frac{1}{c_{train}\left(\bar{f},\bar{e}\right)} \sum_{d \in D} c_d\left(\bar{f},\bar{e}\right) w_i\left(d\right). \tag{5.10}$$

Here, $c_{train}\left(\bar{f},\bar{e}\right)$ is the total count of phrase pair $\left(\bar{f},\bar{e}\right)$ in the training corpus, D is the number of documents in the training corpus, $c_d\left(\bar{f},\bar{e}\right)$ is the count of $\left(\bar{f},\bar{e}\right)$ in document d, and $w_i\left(d\right)$ is the document-level weight of genre feature g_i for document d. Note

Table 5.1: Selection of document-level features inspired by genre-classification literature. The top seven features are most discriminative between the genres NW and UG, and are used in the genre-specific VSM approaches.

Feature
First person pronoun count
Second person pronoun count
Repeating punctuation count ("...", "?!", etc.)
Exclamation mark count
Question mark count
Emoticons count
Numbers count
Third person pronoun count
Plural pronoun count
Average word length
Average sentence length
Total punctuation count
Quote count
Dates count
Percentages count
Long words ($>$ 7 characters) count
Stopwords count
Unique words count

that this definition differs from the standard *tf-idf* weight that is used in Section 5.3.2 since each genre feature has exactly one score per document, and we do not have to normalize for dissimilar subcorpus sizes.

We determine the most genre-discriminating features with a Mann-Whitney U test (Mann and Whitney, 1947) on the observed feature values for each genre in the development set. The seven most discriminative features between the genres NW and UG, which we use in the remainder of this chapter, are shown in the top part of Table 5.1. The main goal of this chapter is to investigate whether this type of genre-revealing features can be useful for the task of translation model genre adaptation, hence we do not attempt to fully exploit the set of possible features. Since genre-discriminating features have the potential to generalize across languages (Petrenz and Webber, 2012), we compute the document-level feature values $w_i\left(d\right)$ on the source as well as the target sides of our bitext, and we examine whether both are equally suitable for translation model genre adaptation.

5.3.4 Genre adaptation with LDA

Next, we use LDA-inferred document distributions as a third vector representation in the adaptation framework. In this formulation, each weight $w_i\left(d\right)$ is a document-level probability for latent dimension i, and weights $w_i\left(\bar{f}, \bar{e}\right)$ are average probabilities of

latent dimension i for all training instances of phrase pair $\left(\bar{f}, \bar{e}\right)$, computed as in (5.10).

We implement LDA using Gensim (Řehůřek and Sojka, 2010), an off-the-shelf tool for LDA inference in Python. By default, Gensim only returns probabilities that are larger than a small threshold α, resulting in incomplete probability distributions. Since we want vectors of equal lengths for all documents in our corpora, we distribute the remaining probability mass uniformly over the missing dimensions. We experiment with varying numbers of latent dimensions (5, 10, 20, and 50), of which LDA with 10 dimensions yields the best translation performance, which is consistent with findings in a related approach by Eidelman et al. (2012). The LDA features in this VSM variant are inferred from the source side of the training data.

5.4 Experimental setup

We evaluate the methods described in Section 5.3 on two Arabic-to-English translation tasks, both comprising NW and UG. The first evaluation set is the Gen&Topic benchmark (see Section 3.3.3), comprising web-crawled news articles and their respective user comments, both covering five different topics. Since this evaluation set has controlled topic distributions per genre, differences in translation quality between genres can be entirely attributed to actual genre differences. The second evaluation set contains NIST OpenMT Arabic-English test sets, using NIST 2006 for tuning, and NIST 2008 and NIST 2009 combined for testing. These data sets cover the genres NW and UG weblogs but are not controlled for topic distributions. Specifications for both evaluation sets are shown in Table 5.2. Note that Gen&Topic contains one reference translation per sentence, while NIST has four sets of reference translations.

Table 5.2: Corpus statistics of the evaluation sets. Numbers of tokens are counted on the Arabic side. Note that Gen&Topic contains one reference translation per sentence, while NIST has four sets of reference translations.

Benchmark			NW	UG	Total
Gen&Topic (1 reference)	Dev	Lines	997	1127	2124
		Tokens	26.9K	25.8K	52.7K
	Test	Lines	1567	1749	3316
		Tokens	46.3K	45.5K	91.8K
NIST (4 references)	Dev	Lines	1033	764	1797
		Tokens	34.4K	14.6K	49.0K
	Test	Lines	1399	1274	2673
		Tokens	46.6K	39.9K	86.6K

All runs use parallel corpora made available for NIST OpenMT 2012, excluding the UN data. While LDC-distributed data sets contain substantial portions of documents within the NW genre, they only contain small portions of UG documents. To alleviate this imbalance we augment our LDC-distributed training data with a variety of web-crawled manually translated documents (see Section 4.3), containing user comments

Table 5.3: Corpus statistics of the Arabic-English parallel training data. Tokens are counted on the Arabic side. Genre mapping: BC=broadcast conversation, BN=broadcast news, NG=newsgroup, NW=newswire, WL=UG weblogs, CM=UG comments, ED=editorials, SP=speech transcripts.

Subcorpus	Genre	Lines	Tokens
NIST broadcast conversation	BC	48K	1,1M
NIST broadcast news	BN	41K	923K
NIST newsgroup	NG	15K	392K
NIST newswire	NW	133K	4.5M
NIST weblog	WL	7.7K	126K
ISI newswire	NW	699K	22.2M
Web newswire	NW	376K	11.1M
Web UG comments	CM	203K	6.0M
Web editorials	ED	127K	4.4M
Web Ted talks	SP	98K	2.2M
Total	All	1.75M	52.9M

that are of a similar nature as the UG documents in the Gen&Topic benchmark, set as well as a number of other genres. Table 5.3 lists the corpus statistics of the training data, split by manual subcorpus labels as used for the subcorpus VSM variant (see Section 5.3.2). While our manually grouped subcorpora approximate those used by Chen et al., exact agreement was impossible to obtain, illustrating that it is not trivial to manually generate optimal subcorpus labels.

We perform our experiments using Oister, see Section 2.4.1 for experimental details. For all experiments, tuning is done separately for the two genre-specific development sets. We use a linearly adapted language model that covers both genres in the benchmark sets, but is not varied between experiments since we want to investigate the effects of different features on translation model adaptation. Finally, we tokenize all Arabic data using MADA.

5.5 Results

In this section we compare a number of variants of the general VSM framework, differing in the way vectors are defined and constructed (see Sections 5.3.2–5.3.4). Statistically significant BLEU improvements are marked with $\triangle$ and ▲ for the $p \leq 0.05$ and the $p \leq 0.01$ level, respectively.

5.5.1 VSM using intrinsic text features

We first test various VSM variants that use automatic indicators of genre and do not depend on the availability of provenance information or manual subcorpus labels (Table 5.4). Of these, genre adaptation with LDA-based features (Section 5.3.4) achieves strongly significant improvements over the unadapted baseline for the NIST-NW and

Table 5.4: BLEU scores of the baseline system and all VSM variants using automatic indicators of genre. Significance is tested against the baseline, and the best performing VSM variant per test set is bold-faced.

		Gen&Topic (1 reference)			NIST (4 references)		
Method		NW	UG	All	NW	UG	All
Baseline		21.5	17.2	19.3	55.3	40.4	48.5
VSM variants using automatic indicators of genre:							
LDA 10 dimensions		21.7 (+0.2)	17.3 (+0.1)	19.4$^{\triangle}$(+0.1)	55.9$^{\blacktriangle}$(+0.6)	40.7$^{\triangle}$(+0.3)	49.0$^{\blacktriangle}$(+0.5)
Genre features	Source	21.9$^{\blacktriangle}$(+0.4)	17.4$^{\triangle}$(+0.2)	19.6$^{\blacktriangle}$(+0.3)	55.7$^{\blacktriangle}$(+0.4)	41.0$^{\blacktriangle}$(+0.6)	49.0$^{\blacktriangle}$(+0.5)
	Target	21.7 (+0.2)	17.5$^{\blacktriangle}$(+0.3)	19.6$^{\blacktriangle}$(+0.3)	55.9$^{\blacktriangle}$(+0.6)	41.2$^{\blacktriangle}$(+0.8)	49.1$^{\blacktriangle}$(+0.6)
Genre+LDA	Source	**21.9$^{\blacktriangle}$(+0.4)**	**17.5$^{\blacktriangle}$(+0.3)**	**19.7$^{\blacktriangle}$(+0.4)**	56.1$^{\blacktriangle}$(+0.8)	41.2$^{\blacktriangle}$(+0.8)	49.2$^{\blacktriangle}$(+0.7)
	Target	21.8$^{\blacktriangle}$(+0.3)	17.5$^{\blacktriangle}$(+0.3)	19.6$^{\blacktriangle}$(+0.3)	**56.2$^{\blacktriangle}$(+0.9)**	**41.2$^{\blacktriangle}$(+0.8)**	**49.2$^{\blacktriangle}$(+0.7)**

Table 5.5: BLEU scores of VSM with manual subcorpus labels in comparison to the best performing VSM with automatic indicators of genre per test corpus (see bold-faced results in Table 5.4), and the combination of manual subcorpus labels and automatic features. BLEU differences and significance for the bottom two variants are measured with respect to VSM manual subcorpora.

	Gen&Topic (1 reference)			NIST (4 references)		
Method	NW	UG	All	NW	UG	All
VSM manual subcorpora	21.6	17.3	19.3	56.3	41.1	49.2
$\triangle$ *wrt unadapted baseline*	*(+0.1)*	*(+0.1)*	*(±0.0)*	*(+1.0)*$^{\blacktriangle}$	*(+0.7)*$^{\blacktriangle}$	*(+0.7)*$^{\blacktriangle}$
VSM automatic genre	21.9$^{\blacktriangle}$(+0.3)	17.5$^{\triangle}$(+0.2)	19.7$^{\blacktriangle}$(+0.4)	56.2 (−0.1)	41.2 (+0.1)	49.2 (±0.0)
VSM manual+automatic	21.9$^{\blacktriangle}$(+0.2)	17.4 (+0.1)	19.6$^{\blacktriangle}$(+0.3)	56.4 (+0.1)	41.4$^{\triangle}$(+0.3)	49.5$^{\blacktriangle}$(+0.3)

the complete NIST test sets, however improvements on the other test portions are very small. When manually inspecting the LDA-inferred latent dimensions, we observe that LDA is overly aggressive in considering all of the UG genre as a single thread, while latent dimensions inferred for NW are more fine-grained. While this finding can be explained by the unbalanced amount of training data per genre, it also illustrates that LDA-based features seem less suitable to capture low-resourced genres.

Next, we evaluate the VSM variant that uses genre-revealing text features inspired by genre classification research (Section 5.3.3). This approach achieves statistically significant improvements over the baseline in all runs except one, i.e., target-side features on Gen&Topic NW. We also see that translation quality is fairly similar for features computed on either side of the bitext, indicating that the proposed genre features can generalize across languages.

Our last VSM variant in Table 5.4 combines genre-revealing and LDA features by using VSM similarities from both approaches as additional decoder features. This combined setting yields the largest improvements, which are all strongly significant and always equal to or better than the performance achieved by either individual feature type, suggesting that the two vector representations are to some extent complementary. Again, source and target genre feature values perform alike, with source-side genre features performing best for Gen&Topic, and target-side genre features obtaining slightly better overall results for NIST.

5.5.2 VSM using manual subcorpus labels

Next we compare our best performing VSM variant per test set (bold-faced in Table 5.4) to the originally proposed VSM variant using manual subcorpus labels (Section 5.3.2). The latter can be considered as an adapted baseline, however with the disadvantage that it relies on the availability of provenance information or manual grouping of documents into informative subcorpora.

Table 5.5 first shows the performance of VSM with manual subcorpus labels, which works well on NIST, confirming previously published results (Chen et al., 2013), but does not lead to significant improvements on Gen&Topic with respect to the unadapted baseline. This suggests that the success of this approach depends on a good fit between the test data distribution and the partitioning of training data into subcorpora, and that a single set of manual subcorpus labels is not guaranteed to generalize to new translation tasks.

The bottom half of the table shows that similar (for NIST) or larger (for Gen&Topic) improvements can be achieved when using the most competitive VSM variant that uses intrinsic text properties instead of manual subcorpus labels. Finally, we use intrinsic text features on top of manual subcorpus labels, i.e., we add all three proposed VSM feature types as additional decoder features. For NIST, this approach yields weakly significant improvements over the runs with only manual subcorpus labels, indicating that the automatic genre features capture additional genre information that is not contained in the manually grouped subcorpora. For Gen&Topic, including manual subcorpus labels does not increase translation performance with respect to VSM with genre and LDA features only, confirming the poor generalization of manual subcorpus labels to new translation tasks.

5.6 Translation consistency analysis

In the proposed translation model adaptation approach, lexical choice is more tailored towards the different genres than in the baseline. We therefore hypothesize that the adapted system increases consistency of output translations within genres. To test this hypothesis, we measure translation consistency following Carpuat and Simard (2012). Their approach studies *repeated phrases*, defined as source phrases p in the phrase table that occur more than once in a single test document d and contain at least one content word. For each repeated phrase, all of its 1-best output translations are compared. If these are identical except for punctuation or stopword differences, the repeated phrase is deemed *consistent*.

The results of the consistency analysis for the unadapted baseline and the best performing VSM genre+LDA variants are shown in Table 5.6. We observe that for both benchmark sets translation consistency is clearly lower in NW than in UG documents. This is likely due to the lower coverage of UG in the training data, which is in agreement with the finding by Carpuat and Simard that translation consistency increases for weaker systems trained on smaller amounts of training data. In line with our expectation, the results also show that document-level translation consistency increases when using the adapted system. Although Carpuat and Simard show that translation consistency does not imply higher quality, they also conclude that consistently translated phrases are more often translated correctly than inconsistently translated phrases.

Table 5.6: Document-level translation consistency values for the baseline and best performing VSM variant using automatic genre indicators.

			% Consistent phrases	
Test set	Genre	# Repeated phrases	Baseline	VSM auto. genre
G&T	NW	7,318	43.2	47.4 (+4.2)
	UG	6,024	55.5	58.2 (+2.7)
	All	13,342	48.7	52.3 (+3.6)
NIST	NW	7,412	40.5	40.6 (+0.1)
	UG	5,431	54.5	57.1 (+2.6)
	All	12,843	46.5	47.6 (+1.1)

Table 5.7 shows some examples of phrases that were translated consistently in one system, but inconsistently in the other. While more phrases moved from being translated inconsistently in the baseline to consistently in the adapted system, the opposite was also observed for all benchmarks. In the UG examples, we see that the adapted system often favors translations that are more colloquial or simplified than (some of) their counterparts in the baseline system, e.g., ‘shows’ instead of ‘indicates’, ‘a year’ instead of ‘annually’, and ‘vaccination’ instead of ‘immunization’. For NW, on the other hand, translations in the adapted system are often more formal, e.g., ‘global’ instead of ‘worldwide’ or more concise, e.g., ‘the health sector’ instead of ‘workers in the health sector’, and ‘east africa’ instead of ‘east african countries’, than in the baseline.

Table 5.7: Examples of source phrases that generate inconsistent translations in the baseline and consistent translations in the adapted system (top), and vice versa (bottom).

Genre	Baseline translation(s)	VSM automatic genre translation(s)
Inconsistent in baseline, consistent in adapted system:		
UG	and this indicates / and this shows that	and this shows
UG	fatigue and stress / and the stress	and the stress
NW	the health sector / workers in the health sector	the health sector
NW	percent of egyptians / percent of them	percent of
Consistent in baseline, inconsistent in adapted system:		
UG	annually	annually / a year
UG	immunization	immunization / vaccination
NW	east african countries	east african countries / east africa
NW	worldwide	worldwide / global

5.7 Conclusions

Domain adaptation is an active field for PBMT, and has resulted in various approaches that adapt system components to specific translation tasks. Motivated by the large translation quality gap that we previously observed between different genres, we have in this chapter focused on genre adaptation. Since most previous approaches depend on the availability of provenance information or manual subcorpus labels, we asked:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

To answer this question, we improved upon an existing adaptation framework by replacing its dependency on subcorpus information with automatic indicators of genre. We explored two types of features that can be automatically extracted from the parallel training and evaluation corpora: genre-revealing features inspired by previous findings in the text genre classification literature, and LDA features.

Experiments on two test sets with two genres showed that the combination of both feature types yields the largest BLEU improvements. This indicates that the proposed genre and LDA features are to some extent complementary. In addition, we observe small improvements over the original adaptation approach when using automatic genre features on top of manual subcorpus labels. We also find that the genre-revealing feature values can be computed on either side of the training bitext, indicating that our proposed features are to some degree language independent. Finally, we find that our genre-adapted translation models encourage document-level translation consistency with respect to the unadapted baseline.

Part II

Translating and Analyzing Informal Language

Research in MT has mostly been driven by *formal* genres, which are characterized by a large degree of standardization and editorial control. Examples of formal genres are news, legal or parliamentary proceedings, and IT or medical guidelines. However, translation of *informal*, typically user-generated (UG), genres containing colloquial language is a more difficult task for PBMT systems. This is reflected by low scores obtained with automatic quality metrics, however little is known about *why* exactly PBMT for informal genres is challenging, and *which aspects* indicate whether translation is easy or not. We address these issues in our second research theme by translating a number of informal text genres and analyzing (i) which errors are common, (ii) why these errors are made, (iii) which aspects can be indicative of poor translation quality, and (iv) what are viable strategies for improving PBMT for informal language.

First, in Chapter 6 we focus a variety of UG sub-genres, which have in common that they have been written by a lay-person, as opposed to a journalist or professional author. To gain a better understanding of the challenges of PBMT for UG text, we conduct a series of analyses on five different UG benchmarks: SMS messages, chat messages, transcripts of telephone conversations, weblogs, and user comments to news articles. We not only contrast informal and formal genres, but also look into the differences between various UG test sets.

Next, an interesting informal genre that has only rarely been used in PBMT research is conversational text. Conversations, or dialogues, involve—by definition—multiple speakers, and are thus noticeably different from formal texts that are typically written by a single writer with a single goal. Different speakers likely have different intentions and language use, possibly demanding PBMT system adaptation at fine-grained levels of dialogue-specific aspects, such as *speakers*, *speaker gender*, *dialogue acts* and *register*. In Chapter 7 we study the impact of these aspects on PBMT quality and their potential to serve as indicators for improved PBMT adaptation.

6 Translating and Analyzing User-Generated Text

6.1 Introduction and research questions

User-generated (UG) text such as found on social media and web forums poses different challenges to SMT than formal text. This is reflected by poor translation quality for informal genres (see for example Figure 6.1), which is typically measured with automatic quality metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), or TER (Snover et al., 2006). These scores alone, however, only reflect the overall translation quality, and do not provide any insights in what exactly makes translating UG text hard. While such knowledge is crucial for improving SMT of UG text, no previous work on error analysis for SMT of user-generated text has been reported.

Moreover, the notion of user-generated content only partially specifies the exact nature of documents. What all documents that can be classified as being UG have in common is the fact that they have been written by a lay-person, as opposed to a journalist or professional author, and that they have not undergone any editorial control. UG text also tends to express the writer's opinion to a larger degree than news articles which generally strive for balance and nuance. Within UG text, we can distinguish several subclasses, including (i) message and dialogue-oriented content such as short message service (SMS) texts, Internet chat messages, and transcripts of conversational speech, (ii) commentaries to news articles, often expressing an opinion about the corresponding articles and relating the content to the reader's situation, and (iii) weblogs, which can bear some resemblance to editorial pieces published by news organizations.

In (Arabic):	قالت عشان العيال متزعلش
Reference:	she said so the kids do not feel upset
MT output:	she said because of the sons
In (Chinese):	你 路上 慢 点
Reference:	take your time
MT output:	you are on the road to slow points

Figure 6.1: SMS examples with poor PBMT output.

While UG text processing tasks are becoming more common, most research in MT is still driven by formal translation tasks,[1] and existing error analysis approaches may only be partially useful for UG. To explain the typically poor PBMT performance observed for UG texts, we conduct a series of analyses on five UG benchmarks for two language pairs, Arabic-English and Chinese-English, by asking:

RQ4 *How is translation quality of PBMT influenced by different types of user-generated text?*

To answer this question, we compare our five UG benchmarks both quantitatively and qualitatively. Our quantitive analysis starts with measuring translation quality, average translation phrase length, and model coverage of source phrases, target phrases and phrase pairs of various lengths. Next, we also compare for each UG test set the distribution of errors caused by (i) unseen source words, (ii) unseen target words, or (iii) suboptimal scoring of possible translation candidates. Our qualitative analysis comprises sentence-level annotations representing these errors. Together, these analyses answer the following question:

RQ4a. *What are the most common error types for various UG genres?*

In addition, we also run all of our analyses for two news benchmarks, allowing us to compare PBMT performance and errors across UG test sets and between UG text and more formal text. Doing so, we address the following question:

RQ4b. *To what extent do PBMT errors differ for different types of UG text and between UG text and news?*

Finally, we want to identify translation modeling aspects that should be addressed to improve translation of UG data, and we therefore ask:

RQ4c. *What are promising strategies to improve PBMT for UG genres?*

Organization This chapter is organized as follows: In Section 6.2 we discuss relevant related work. In Section 6.3, we briefly discuss the experimental setup and the data sets used in this chapter. Next, in Section 6.4 we describe all of our quantitative and qualitative analyses and discuss the results and findings. Finally, we provide conclusions in Section 6.5.

6.2 Related Work

Relevant previous work for this chapter concerns two research subjects: error analysis for PBMT and PBMT for informal language, both of which we discuss below.

[1] One of the few exceptions is NIST OpenMT 2015, which focusses entirely on translating UG genres.

6.2.1 Analyzing PBMT errors

Identifying and analyzing different types of PBMT errors is an essential step towards the development of translation approaches that can achieve more robust performance, and has therefore been the focus of earlier work. The majority of previous work on PBMT error analysis has dealt with categorizing errors and designing error taxonomies, typically for post-editing purposes. Such taxonomies often use part-of-speech (POS) tags or other syntactic or linguistic information to classify PBMT output errors into one of several categories, reflecting various granularity levels of inflectional errors, reordering errors, lexical errors, and punctuation errors (Font Llitjós et al., 2005; Vilar et al., 2006; Bojar, 2011; Popović and Ney, 2011). Giménez and Màrquez (2008, 2010) use a large set of existing automatic evaluation metrics to assess PBMT errors. Other work considers very specific aspects of PBMT output errors, such as fluency (Elliott et al., 2004) or verb inflections (Popović and Ney, 2006). Finally, Costa et al. (2015) present a comprehensive taxonomy covering most of the classes considered in previous work.

All of the above have in common that they only look at PBMT output, and do not provide insights into *why* certain mistakes are made during translation. In addition, the presented analyses often deal with language-specific phenomena, for example for translating into Chinese (Vilar et al., 2006), Czech (Bojar, 2011), or Spanish (Popović and Ney, 2006). One of the few works analyzing the causes of PBMT errors rather than their appearance is performed by Irvine et al. (2013a), who use word alignment links to determine whether incorrect lexical choices are made due to poor source word coverage, poor target word coverage, suboptimal scoring of translation options, or search errors during decoding. They mainly use their error analysis to determine how the observed errors change when shifting domains, however it can be applied to any translation task in any language pair.

Finally, a few other efforts have analyzed PBMT in the context of domain adaptation, for example by examining at which stage of the PBMT pipeline the available in-domain data can best be used (Duh et al., 2010), how in-domain and out-of-domain models can best be combined (Bisazza et al., 2011), or whether it is more promising to improve either phrase extraction or scoring (Haddow and Koehn, 2012).

6.2.2 PBMT for informal language

The vast majority of PBMT research, including the above described work on error analysis, is evaluated on test sets containing relatively formal language, such as news, legal or parliamentary proceedings, and IT or medical guidelines. Work on PBMT of informal, UG text mostly targets reduction of OOV words in the source text, for example by correcting spelling errors (Bertoldi et al., 2010), normalizing noisy text to more standardized text (Banerjee et al., 2012; Ling et al., 2013a), or enhancing the training data with bilingual segments extracted from Twitter (Jehl et al., 2012; Ling et al., 2013b). Other work improves PBMT of UG text by combining statistical and rule-based MT (Carrera et al., 2009), or models trained on more formal and less formal genres (Banerjee et al., 2011). Finally, Roturier and Bensadoun (2011) conduct a comparative study to determine the ability of several PBMT systems to translate UG text, but they do

not examine what errors the systems make. To our knowledge, our work is the first that looks inside an PBMT system to systematically inspect its behavior across a diverse spectrum of UG text types.

6.3 Experimental setup

We perform our error analysis on five UG data sets in two language pairs, Arabic-English and Chinese-English. Below we describe the details of our data and experimental setup.

6.3.1 Evaluation sets

For both language pairs we use evaluation sets for five types of user-generated text: SMS messages, chat messages, manual transcripts of phone conversations (called Conversational Telephone Speech (CTS)), weblogs, and readers' comments to news articles. The first four data sets originate from BOLT and NIST OpenMT, and are distributed by the Linguistic Data Consortium (LDC), while the last data set is crawled from the web (see Section 4.3 for details on the harvesting procedure). All UG experiments are contrasted with two news data sets: the news portions of NIST evaluation sets, and web-crawled news articles.

For Arabic-English, the web-crawled news articles and comments originate from the Gen&Topic data set (introduced in Section 3.3.3), in which both genres cover the same distributions over various topics. Consequently, any observed differences between the news and UG portions of this data set can be entirely attributed to genre differences and not to potential topical variation.

We have attempted to create similar-sized and reasonably large benchmark sets. However, benchmarks for some genres are smaller since data is for these genres is scarce. Tables 6.1 and 6.2 show the data specifications of the Arabic-English and Chinese-English evaluation sets, respectively. Note that two of the NIST evaluation sets contain four reference translations instead of one. To allow for a fair comparison, we compute in all our analyses scores for these benchmarks using a single reference translation and average the scores of the four references.

6.3.2 PBMT systems

All experiments presented in this chapter are performed with our in-house phrase-based PBMT system Oister (see Section 2.4.1). Our Arabic-English system is built from 1.75M lines (52.9M source tokens) of parallel text, and our Chinese-English system from 3.13M lines (55.4M source tokens) of parallel text. We tokenize all Arabic data using MADA (Habash and Rambow, 2005), ATB scheme, and we segment the Chinese data following Tseng et al. (2005). Both systems use an adapted 5-gram English language model that linearly interpolates different English Gigaword subcorpora with the English side of our bitexts, containing both news and UG data.

While parallel data is scarce in general, the situation is much worse for UG data, where there are hardly any sizable parallel corpora for any language pair. As a consequence, the training data of both systems comprises 70–75% news data, mostly

Table 6.1: Statistics of the Arabic-English UG (top) and contrastive news (bottom) evaluation sets. Tokens are counted on the Arabic side.

	Dev set		Test set		
Genre	Lines	Tokens	Lines	Tokens	Refs
SMS	2.7K	23.3K	7.6K	44.9K	1
Chat	3.5K	22.5K	7.1K	44.5K	1
CTS	2.4K	23.1K	3.6K	40.6K	1
Comments	1.1K	25.8K	1.7K	45.5K	1
Weblogs	0.8K	14.6K	1.3K	39.9K	4
News 1	1.0K	26.9K	1.6K	46.3K	1
News 2	1.0K	34.4K	1.4K	46.6K	4

Table 6.2: Statistics of the Chinese-English UG (top) and contrastive news (bottom) evaluation sets. Tokens are counted on the Chinese side.

	Dev set		Test set		
Genre	Lines	Tokens	Lines	Tokens	Refs
SMS	1.8K	15.3K	4.2K	36.3K	1
Chat	4.0K	25.6K	6.0K	45.7K	1
CTS	2.2K	25.1K	2.9K	44.8K	1
Comments	1.0K	26.5K	1.5K	41.0K	1
Weblogs	0.5K	8.8K	0.7K	14.4K	4
News 1	0.8K	24.5K	1.5K	41.9K	1
News 2	1.2K	29.4K	0.7K	17.7K	4

LDC-distributed, and 25–30% data in various other genres (weblogs, comments, editorials, speech transcripts, and small amounts of chat data), mostly harvested from the web. Per language pair, all experiments use the same PBMT models, but we tune parameters separately for each benchmark.

To put the results of our system into perspective, we also run a first series of experiments on a well-known and established online PBMT system.

6.4 Error analysis and results

In this section, we perform four different experiments series on all of our benchmarks. First, we quantify the gap in translation quality between news and different types of UG data (§6.4.1). Next, we investigate what kind of translation choices the PBMT system makes for the different benchmarks (§6.4.2), and what kind of translation choices the PBMT system could have made (§6.4.3). Finally, we examine why the PBMT system made the translation choices that it made (§6.4.4).

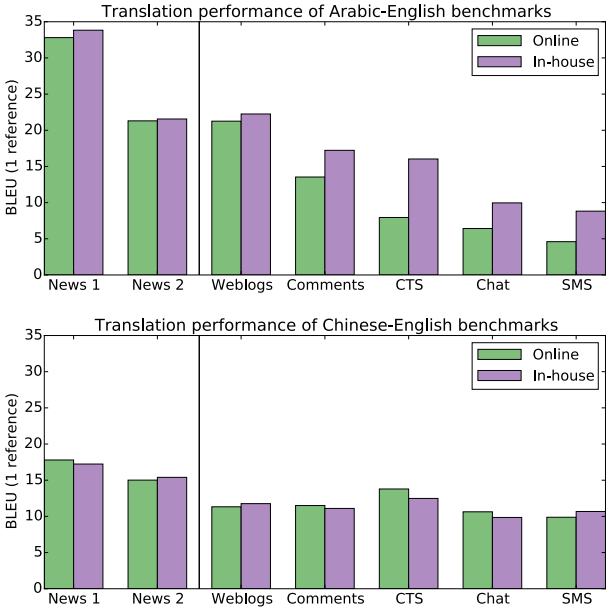


Figure 6.2: Translation performance of baseline experiments for two news and five UG Arabic-English (top) and Chinese-English (bottom) benchmarks, measured in case-insensitive BLEU for one reference translation.

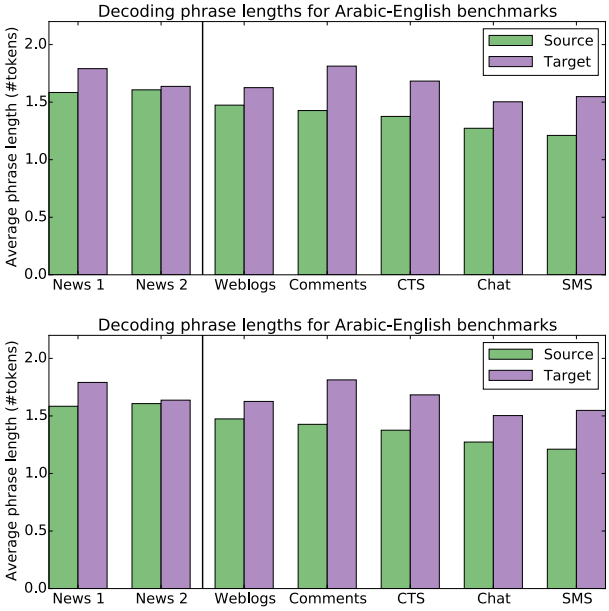


Figure 6.3: Average source-side and target-side phrase lengths used during decoding of two news and five UG Arabic-English (top) and Chinese-English (bottom) benchmarks.

6.4.1 Overall translation quality

A first important indication of PBMT quality across different genres can be given by translation quality measures that are based on the similarity between the PBMT output and a reference human translation. To estimate the gap in translation quality between news and UG text, but also among various types of UG text, we measure the BLEU scores (1 reference) of our in-house PBMT system and that of the online system on all our evaluation sets.

The results in Figure 6.2 (top) show that translation quality differs greatly between the Arabic-English data sets. In particular, the News 1 data set (from NIST) yields considerably higher BLEU scores than all other evaluation sets, including the News 2 (from Gen&Topic) set, which represents the same genre but is visibly more difficult to translate. On the other end of the spectrum, we see that translation quality of the SMS and chat data sets is very poor. Note that our in-house system is optimized per genre, whereas the online system is optimized for general language and speed.

For Chinese-English (Figure 6.2, bottom) the differences in BLEU are less pronounced, both across the different data sets and between the two PBMT systems. Still, translation quality is worse for the UG data sets than for news, indicating that also for this language pair translating UG text is more challenging than translating news.

As all subsequent analyses require system-internal information, we carry out the experiments with our in-house system only.

6.4.2 Translation phrase length analysis

Most state-of-the-art PBMT systems, including our in-house system, are phrase-based, with translations being generated phrase by phrase rather than word by word (Koehn et al., 2003). An abundant usage of small phrases during decoding indicates that the system is not taking advantage of the model's ability to memorize large contextual and possibly non-compositional translation blocks. It is therefore interesting to measure the average phrase length, i.e., number of tokens, used by the system, for the source as well as the target language (Figure 6.3). For Arabic-English (top) we see that source-side phrases are noticeably longer for both news benchmarks than for the UG data sets. The average target-side phrase length, on the other hand, shows less correlation with the genres of the data sets. Similar trends are observed for Chinese-English (bottom), however differences are less extreme.

In general, PBMT systems incur higher model costs when utilizing many small phrases rather than few large phrases. If, in spite of that, a system selects many short phrases, which is the case for most of our UG benchmarks, this can be due to (i) unreliable translation probabilities or (ii) to the mere lack of correct translation options in the models. We investigate both issues in the following analyses.

6.4.3 Model coverage analysis

Next, we examine the translation model coverage for each data set, which tells us what phrases the system *could have* used for decoding. For each of our test sets, we create automatic word alignments using GIZA++ (Och and Ney, 2003), and extract from these

Table 6.3: Target language model perplexity and translation model coverage of Arabic-English benchmarks. Phrase pair recall values are broken down by source phrase length. Intensities of the cell colors indicate relative recall values with respect to the best scoring benchmark (measured in BLEU).

			Source phrase recall				Target phrase recall				Phrase pair recall			
Genre	BLEU	LM PP	1	2	3	4	1	2	3	4	1	2	3	4
News 1	33.8	65	99.7	88.9	56.3	26.1	99.7	91.1	61.5	29.6	84.9	54.4	23.6	8.1
News 2	21.5	86	99.6	88.1	53.7	21.8	99.5	88.1	53.4	23.6	77.4	46.9	18.8	5.9
Weblogs	22.3	152	99.2	80.5	40.6	13.5	99.5	86.3	48.9	17.8	78.4	41.5	12.9	2.9
Comments	17.2	117	97.7	80.2	43.0	15.3	99.7	89.8	55.3	21.9	59.1	33.2	11.1	2.8
CTS	16.0	103	97.4	66.3	25.1	6.4	99.8	90.8	54.3	21.5	66.7	25.7	6.1	1.0
Chat	10.0	179	94.1	56.0	19.4	4.7	98.6	86.1	47.3	16.7	60.8	21.3	4.5	0.8
SMS	8.8	196	93.7	57.8	17.5	3.3	99.1	86.3	47.0	14.6	62.0	21.1	3.7	0.4

Table 6.4: Target language model perplexity and translation model coverage of Chinese-English benchmarks. See Table 6.3 for detailed explanation.

			Source phrase recall				Target phrase recall				Phrase pair recall			
Genre	BLEU	LM PP	1	2	3	4	1	2	3	4	1	2	3	4
News 1	17.2	121	99.0	80.2	40.8	16.2	99.5	84.9	48.0	19.5	69.1	34.8	10.8	3.3
News 2	15.4	118	98.8	84.2	44.3	16.0	99.4	83.8	44.2	14.7	63.1	32.4	10.7	3.3
Weblogs	11.8	153	98.6	76.6	33.8	11.1	99.3	81.6	40.8	12.4	59.0	27.0	7.3	1.7
Comments	11.1	195	98.7	78.3	35.2	8.7	97.9	77.9	35.1	10.2	53.5	21.6	5.0	1.0
CTS	12.5	135	98.7	80.7	40.1	10.5	99.8	86.3	47.4	16.4	70.0	33.5	9.3	1.7
Chat	9.9	221	98.0	71.9	27.5	6.1	99.4	82.6	43.2	13.0	62.3	24.8	5.4	0.6
SMS	10.7	234	97.3	68.5	24.9	4.8	99.0	80.4	40.5	12.5	62.6	24.6	5.1	0.5

the set of all reference phrase pairs using Moses' phrase extraction algorithm (Koehn et al., 2007). By comparing this set of phrase pairs to the available phrases in the PBMT models, which have been extracted using the same procedure, we can compute the following statistics:

1. *Source phrase recall*, defined as the fraction of reference phrase pairs whose *source* side is found in the PBMT models.
2. *Target phrase recall*, defined as the fraction of reference phrase pairs whose *target* side is found in the PBMT models.
3. *Phrase pair recall*, defined as the fraction of reference phrase pairs whose source and target side are jointly found in the PBMT models.

Low recall values indicate that the models lack phrases or phrase pairs that match the test data, which can be addressed by adding additional relevant training data or by generating new phrases. In addition, we measure language model perplexity as an indication of how predictable each benchmark is for the language model: high perplexity corresponds to lower coverage. Note that this is an extended version of our coverage analysis presented in Section 3.4.2, which excludes target phrase recall. Formal mathematical definitions of the above statistics are also provided in Section 3.4.2.

The model coverage results for Arabic-English and Chinese-English are shown in Tables 6.3 and 6.4, respectively. All recall scores are broken down by phrase length, up to phrases of four tokens. The source-target phrase pair recall (last four columns) is split by source phrase length rather than target phrase length since source phrases are the actual input to the PBMT system. We use shaded cells of which the intensity represents relative recall values with respect to the best scoring benchmark according to BLEU, i.e., News 1.

The results show that source phrase recall is substantially lower for the UG benchmarks than for news, particularly for longer phrases. Regarding target phrase recall, differences between various data sets and genres are much smaller. This suggests that many of the reference phrases could potentially be generated by the system, even for the UG data. However, to be able to output the available target phrases, the system needs a match with the input source phrases, which is exactly what is being measured with phrase pair recall. Here, we see that for the majority of single-word source phrases, the expected target phrase is accessible by the system. For longer phrases, though, there is again a drastic decline in recall, with almost no phrases of length 4 or longer having the expected target covered by the models. Similar to source phrase recall, this decline is notably bigger for UG than for news.

Looking at the differences between the various types of UG data, we see that the SMS and chat benchmarks are most severely affected by overall poor model coverage. As for weblogs, the target phrase recall is similar to SMS and chat, whereas both source phrase and phrase pair recall are much higher. For CTS and web comments, there are notable differences between model coverage for the two language pairs, despite similar BLEU scores. While comments have better coverage in the Arabic-English models, CTS has higher recall values for Chinese-English.

Finally, we see that language model perplexity is on average lower for Arabic-English than for the Chinese-English benchmarks. This is somewhat surprising given

that perplexity is measured on the English side, but it can partially explain the low BLEU scores on, for example, the Chinese-English News 1 benchmark. All news benchmarks have relatively low perplexities, which is expected since the language model covers more news than UG data. Of the UG benchmarks, CTS has a remarkably low perplexity value, suggesting that for this genre the language model can potentially compensate for low translation model coverage.

6.4.4 WADE: Word Alignment Driven Evaluation

To gain a more fine-grained insight in *why* our PBMT system makes its translation choices, we reimplement an evaluation approach proposed by Irvine et al. (2013a), which analyzes PBMT error types at the word alignment level. The analysis exploits automatic word alignments between (i) a given source sentence and its reference translation, and (ii) the same source sentence and its automatic translation. Each aligned source-reference word pair is examined for whether the alignment link is matched by the decoder. Formally, f_i is a foreign word, e_j is a reference word aligned to f_i, $a_{i,j}$ is the alignment link between f_i and e_j, and H_i is the set of output words that are aligned to f_i by the decoder. If $e_j \in H_i$, the alignment link $a_{i,j}$ is marked as correct. Otherwise, $a_{i,j}$ is categorized with one of the following error types:

1. A SEEN error indicates an unseen source word, i.e., out-of-vocabulary (OOV) item. This error is assigned to $a_{i,j}$ if f_i does not appear in the phrase table used for translation. This type of error inversely correlates with length-1 source phrase recall (§6.4.3).

2. A SENSE error indicates an unseen target word. This error is assigned to $a_{i,j}$ if f_i does appear in the phrase table but never with translation candidate e_j.

3. A SCORE error indicates suboptimal scoring of translation options. This error is assigned to $a_{i,j}$ if f_i exists in the phrase table with translation candidate e_j, but another translation candidate is preferred by the decoder.

Irvine et al. consider a final category of *freebies*: OOVs that are copied over verbatim to the output sentence and accidentally match the reference translation, e.g., urls, proper nouns, etc. For the language pairs that we study, they are very rare due to the differences in writing scripts: at most 0.35% for Arabic-English (in CTS) and 0.63% for Chinese-English (in SMS). Manual inspection reveals that nearly all freebies are English words in the foreign source text. Since they are so rare, we omit freebies from our results.

As WADE errors are assigned at the fine-grained level of individual words, this analysis allows for (i) sentence-level visualization of errors, and (ii) collecting aggregate statistics of each error type for an entire evaluation set. By assembling the latter for various benchmarks, we can quantify global differences between genres or data sets (see quantitative analysis below). At the same time, by examining (i) we can gain insight in the nature of the different 'errors', which might be real mistakes, or, for instance, different lexical choices (see qualitative analysis below).

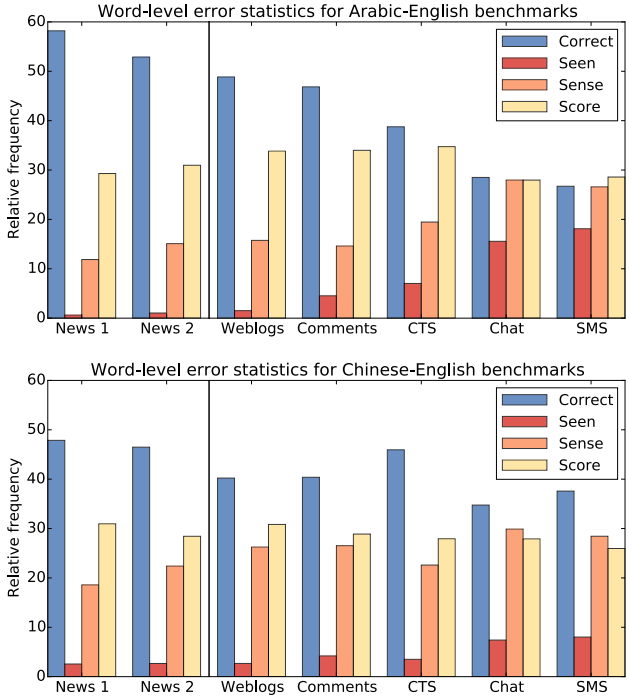


Figure 6.4: Aggregate WADE error statistics for Arabic-English (top) and Chinese-English (bottom) benchmark sets.

Quantitative results

The aggregate error statistics for each data set are shown in Figure 6.4. To put our results into perspective, we recall the findings of Irvine et al. (2013a). They find that for *formal* domains using a French-English system, 50–60% of the alignment links are correct, and SCORE errors are more common than SENSE errors, which in turn are more common than SEEN errors. While we observe a similar distribution for our Arabic-English news benchmarks, these numbers do not generalize to the Arabic-English UG benchmarks nor to any of the Chinese-English data sets.

First, the portion of SEEN errors increases dramatically for the Arabic-English UG translation tasks. For Chinese-English this trend is less pronounced yet also clearly observable. Next, SENSE errors also increase substantially for most of the UG data, making up the majority of the errors for Chinese-English SMS and chat. This indicates that a promising strategy for adapting PBMT systems to translating UG data involves generating new target-side translation candidates that match the source phrases in the input sentences. Finally, we evaluate the fraction of SCORE errors. While this is the most commonly observed error type in most of the data sets, there seems to be very little correspondance with the genre or BLEU scores of the benchmarks. This is an interesting finding since most work in system adaptation for PBMT focuses on better

scoring of existing translation candidates (Matsoukas et al., 2009; Foster et al., 2010; Axelrod et al., 2011; Chen et al., 2013, among others). However, for UG translation tasks this does not appear as the most effective approach.

Qualitative results

The generated sentence-level error annotations allow us to examine the various error types in detail. The first phenomenon that we repeatedly observe in the UG data are SEEN errors due to misspellings or, in the case of Arabic, dialectal forms. Two such examples are shown in Figures 6.5A and 6.5B: In the first, the PBMT system does not recognize the dialectal form of verb negation 'mtzEl$',[2] which is a morphologically complex word containing both a prefix and a suffix. In the second, the input word 'Almw**b**Ayl' ('mobile') is wrongly spelled as 'Almw**y**Ayl'. It is interesting to note that 'b' and 'y' are very similar in the Arabic script. This type of errors is particularly frequent in chat and SMS, which can partly explain the different distribution of errors across the Arabic-English data sets (Figure 6.4).

Also frequently observed in the UG data are PBMT lexical choices that are more formal than the reference translations. We suspect that this is caused by the large amount of formal data in the PBMT models, and it stresses the need for adaptation to UG data. Often, the optimal lexical choice is simply absent from the PBMT models, resulting in SENSE errors. This can be observed in Figure 6.5A, where 'sons' is output instead of 'kids', and in Figure 6.5C, where 'i understand' is output instead of the colloquial 'i got it'. In other situations, the annotated SCORE errors indicate that the correct choice was available to the PBMT system without being selected for translation. For example in Figure 6.5D, the output 'my parents' is preferred to the more colloquial 'mom and dad' in the reference.

Another phenomenon, particularly common for Chinese-English UG translations, is that idioms are translated in small chunks, thereby losing their meaning as a phrase. In Figure 6.5D, the characters '说', '一', and '声' mean 'to say', 'one', and 'sound', respectively. The phrase '说一声' as a whole means 'talk a bit about something' but is not covered by the PBMT models. Similarly, '你路上慢点' in Figure 6.5E literally means 'you on the road slow a bit', which, if covered by the models, could have been translated into 'be careful on your way' or 'take your time'. These examples illustrate that the low phrase pair recall for longer phrases severely complicates PBMT of UG data.

A final recurring issue in SMS and chat messages is the omission of first person pronouns, see for example Figure 6.5E. The Chinese source phrase '上网了' literally means 'get online' (+ auxiliary word marking past tense). A native speaker understands that this concerns the sender, which is reflected by a first person pronoun in the reference. The PBMT system, on the other hand, cannot infer the subject of this phrase and instead generates a translation without pronouns.

Other, less common, types of errors occurring in the UG data are due to inconsistent segmentation or tokenization of input text, which mostly affects rare words, emoticons, and repeating punctation. Finally, SEEN errors for named entities are overall rare but occur in both news and UG benchmarks.

[2] We use Buckwalter transliteration to represent Arabic text in a left-to-right orientation.

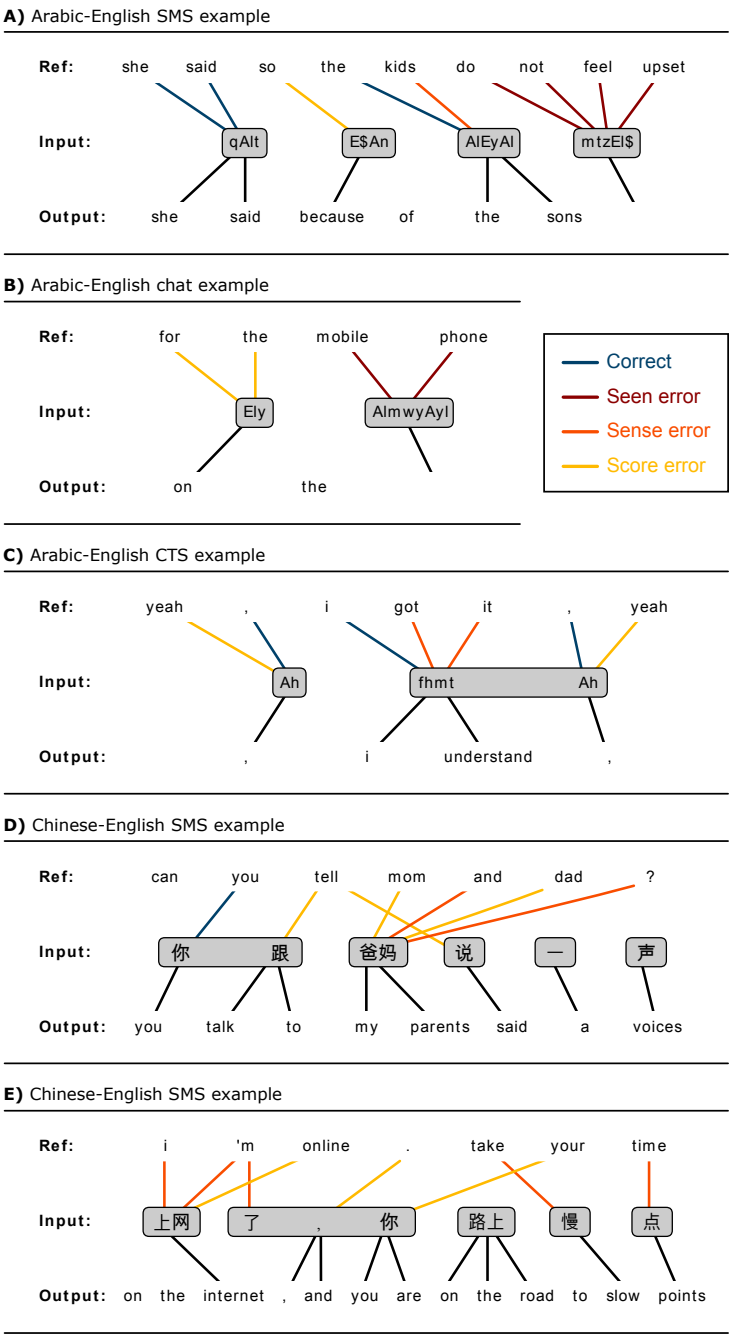


Figure 6.5: Sentence-level error annotations from various UG benchmarks illustrating common issues in PBMT of UG data. We use Buckwalter transliteration to represent the Arabic source text in a left-to-right orientation.

6.5 Conclusions

Translating user-generated (UG) text such as found on social media and web forums is a difficult task for PBMT. This is reflected by poor translation quality, which is typically measured with automatic quality metrics such as BLEU, METEOR, or TER. These scores alone, however, do not provide any insights in what exactly makes translating UG text hard. We therefore performed in this chapter an extensive error analysis explaining the poor translation quality often observed for UG text.

Since the notion of user-generated content only partially specifies the exact nature of documents, we have collected and used five different types of UG data: SMS messages, chat messages, manual transcripts of phone conversations, weblogs, and readers' comments to news articles. What all these documents have in common is the fact that they have been written (or spoken) by a lay-person, as opposed to a journalist or professional author, and that they have not undergone any editorial control. However, different types of UG text can have very different characteristics and pose different challenges to PBMT, which we addressed by asking:

RQ4 *How is translation quality of PBMT influenced by different types of user-generated text?*

To answer this question, we compared our five UG benchmarks both quantitatively and qualitatively. Our quantitative results show among others that (i) UG data is translated with shorter source phrases than news, (ii) UG translation model coverage deteriorates substantially for longer phrases, and (iii) phrase-pair OOVs pose a bigger challenge to UG translation tasks than source OOVs. In our qualitative analysis we found that common issues in UG data include (i) OOVs due to misspellings or Arabic dialectal forms, (ii) lexical choices that do not reflect colloquial formulations, (iii) phrasal idioms being translated word by word, and (iv) omitted first person pronouns in SMS and chat.

In addition, when comparing to two news benchmarks, we found that the SMS and chat benchmarks are the most distant from formal text at all the analyzed levels. Errors in other types of UG are often more similar to news errors than to those in SMS and chat messages. Since we found that different types of UG exhibit such dissimilar error distributions, they demand diverse strategies to improve PBMT quality. For example, SMS and chat data might benefit from text normalization (Bertoldi et al., 2010; Yvon, 2010; Ling et al., 2013a) or otherwise resolving source OOVs, while some of the other UG genres likely benefit from better scoring of existing translation candidates.

7

Translating and Analyzing Fictional Dialogues

7.1 Introduction and research questions

Research in PBMT has mostly been driven by translation tasks using relatively formal text such as parliamentary proceedings. However, as shown in Chapter 6, PBMT of informal genres is much more challenging, and dialogue-oriented content such as SMS and chat messages are particularly problematic. We therefore focus in this chapter on PBMT for dialogues, an informal genre that involves, by definition, multiple speakers, and is thus noticeably different from other types of text (Fernández, 2014). Other, more formal text is typically written by a single, often professional, writer with a clear intention, e.g., informing or persuading, and moreover has been editorially controlled according to standards of language use. In dialogues, on the other hand, different *speakers* have different intentions and language use. Such variations are reflected *dialogue acts*, functional actions such as questions or answers (Bunt, 1979), and by *register*, a term referring to socio-situational language variation (Lee, 2001), respectively.

While these and other dialogue-specific aspects have been analyzed in dialogue research (Schlangen, 2005; Fernández, 2014), their impact on PBMT has hardly been studied. In this chapter, we take a first step towards investigating the effect of dialogue acts, speakers, speaker gender, and register on PBMT performance by asking:

RQ5 *What impact do dialogue-specific aspects have on translation quality?*

A likely explanation for the fact that PBMT for dialogues has hardly been addressed before is the lack of adequate evaluation data, i.e., parallel conversations annotated with dialogue-specific variables. To overcome this limitation we first create and release a corpus of multilingual movie dialogues annotated with the four aforementioned dialogue-specific aspects. Since movie dialogues are fictional, we can only consider them as an approximation of real face-to-face conversation. However, several corpus-based studies have shown that, while movie dialogues differ from natural spoken dialogues in terms of spontaneity—they exhibit fewer incomplete utterances, hesitations, and repetitions—, they do not differ to a great extent in terms of linguistic features

and main messages (Forchini, 2009, 2012; Dose, 2013). In fact, movie dialogues approximate real face-to-face conversation more than for example SMS and chat, in which media constraints influence the flow of a conversation (Whittaker, 2003; Brennan and Lockridge, 2006). Finally, Danescu-Niculescu-Mizil and Lee (2011) have shown that certain psycholinguistic and gender-specific aspects of language are also observed in fictional dialogues, indicating that conclusions drawn from experiments on fictional dialogues generalize at least partially to real spoken conversations.

Armed with our newly annotated dialogues, we can measure PBMT quality fluctuations among different values of each dialogue aspect, thus asking:

RQ5a. *To what extent does PBMT quality vary between different speakers, speaker genders, dialogue acts and register?*

Concretely, we use an approximate randomization approach to measure whether BLEU fluctuations for each dialogue aspect are larger than expected at random. If so, this would indicate that we may be able to (i) predict PBMT output quality by having access to dialogue aspect information, and (ii) successfully apply PBMT adaptation at fine-grained level of various dialogue aspects. To verify the former hypothesis, we ask:

RQ5b. *Which dialogue aspects are most indicative for PBMT quality?*

To verify the second hypothesis, we ask:

RQ5c. *Can automatic annotations of dialogue-specific aspects benefit adaptation for PBMT?*

To this end, we perform a number of preliminary adaptation experiments towards two dialogue aspects for which values can be heuristically detected from the training corpus, namely dialogue acts and registers.

Organization. This chapter is organized as follows: first, in Section 7.2, we discuss related work. Next, in Section 7.3 we describe our efforts to automatically annotate multilingual movie dialogues with four dialogue-specific variables. In Section 7.4 we present our approach to measuring how PBMT quality is affected by each of the four dialogue-specific aspects, and we discuss and analyze the results in Section 7.5. Next, in Section 7.6 we investigate the hypothesis that PBMT of fictional dialogues benefits from adaptation towards various dialogue acts and registers. Finally, we summarize our findings in Section 7.7.

7.2 Related work

While previous work on PBMT for dialogues is not abundant, a few lines of related work are relevant for this chapter. Below we discuss previous work on annotating fictional dialogue corpora and research studying dialogue and discourse phenomena in the context of PBMT.

7.2.1 Aligning and annotating fictional dialogue corpora

Automatic alignment or annotation of fictional dialogues has resulted in a number of corpora, the most well-known being the OpenSubtitles corpus (Tiedemann, 2009a; Lison and Tiedemann, 2016). This corpus contains non-professionally translated subtitles of movies and TV series, collected from `www.opensubtitles.org` and cross-lingually aligned using time information, bilingual lexicons, and cognates (Tiedemann, 2008). In recent work, Tiedemann (2016) exploits the fact that often multiple subtitle versions of a single movie exist, which can be used to create multi-reference corpora yielding small BLEU improvements.

OpenSubtitles corpora contain noisy bitexts due to their non-professional translations and automatic alignment, and therefore yield variable PBMT performance. Better PBMT quality can be achieved when training a system on professionally high-quality subtitle translations (Volk, 2009; Volk et al., 2010; Petukhova et al., 2012). Unfortunately, such corpora are not publicly available.

A different source of fictional dialogues is the Internet Movie Script Database (IMSDb),[1] containing movie scripts with speakers, utterances, and context, e.g., change of scenes. Unlike subtitles, IMSDb scripts are not multilingual and only cover the English language. However, the available speaker names provide interesting additional information, which has been exploited in previous work. For example, Danescu-Niculescu-Mizil and Lee (2011) use speaker and gender annotations to show that the psycholinguistic 'chameleon effect'[2] is also observed in fictional dialogues. Besides speaker names and gender information, Walker et al. (2012) also annotate movie dialogues with detailed character descriptions. The resulting corpus is used to create character models that can be applied to character-specific language generation. Banchs (2012) also presents an annotated English movie dialogue corpus, with a main focus on its potential usability to train chat-oriented dialogue systems.

In this chapter we combine subtitles and scripts to annotate multilingual utterances with speaker and other dialogue information. Recently, some related work has described similar efforts (Wang et al., 2016; Lison and Meena, 2016; Bawden et al., 2016), indicating that there is an increasing interest in multilingual annotated conversational corpora. While Lison and Meena (2016) use their speaker-annotated utterances in a non-MT scenario, both Wang et al. (2016) and Bawden et al. (2016) show, in a small-scale setting, that PBMT adaptation to speaker genders improves BLEU. This confirms our hypothesis that dialogue aspects such as speaker gender can be exploited to improve PBMT quality. In this chapter we show that not only speaker gender, but also other dialogue aspects can be used as indicators for PBMT adaptation.

7.2.2 Dialogue and discourse in PBMT

While PBMT for dialogues has rarely been addressed, there is a growing body of work on discourse in PBMT, which aims to improve grammatical coherence of output translations. Discourse in PBMT typically involves translation and prediction of discourse connectives (e.g., but, because) or pronouns (e.g., it, he, she). Work on the former

[1] `www.imsdb.com`

[2] The effect that people adjust their language to their conversation partner.

includes disambiguating discourse connectives (Meyer et al., 2012) or analyzing the impact of discourse structure on MT (Meyer and Webber, 2013; Li et al., 2014). Work on cross-lingual pronoun prediction has largely been driven by shared tasks of the Workshop on Discourse in Machine Translation (DiscoMT, Hardmeier et al. (2015); Guillou et al. (2016)), and includes predicting correct translations of pronouns when translating into morphologically richer languages (Hardmeier et al., 2013; Hardmeier, 2014) or distinguishing different pronoun functionalities (Guillou, 2016). A recently addressed discourse task concerns improving English question tags (e.g., isn't it?, are you?), which are particularly frequent when translating into English (Bawden, 2017).

Since we are in this chapter considering PBMT of various dialogue acts, including questions, our work also touches upon previous work on cross-lingual question answering, in which either the question or the retrieved document needs to be translated into another language. Most work in this field uses MT only as a tool to translate questions or documents into English (Giampiccolo et al., 2007; Forner et al., 2008). Exceptions focusing on optimal, task-specific, translations are made in work by Tiedemann (2009b) and Ture and Boschee (2016).

7.3 Corpus construction and annotation

To measure the effect of dialogue acts, speakers, speaker gender, and register on PBMT performance we need a multilingual dialogue corpus in which utterances are annotated with each of these dialogue aspects. Unfortunately, existing corpora are limited to the English language (Janin et al., 2003; McCowan et al., 2005; Danescu-Niculescu-Mizil and Lee, 2011; Banchs, 2012; Walker et al., 2012) or contain only some of the required annotations (Wang et al., 2016). We therefore first automatically annotate multilingual movie dialogues with the above dialogue dimensions for five language pairs: Arabic-English, Chinese-English, Dutch-English, German-English, and Spanish-English.

7.3.1 Annotation process

For our annotation process we build on OpenSubtitles, IMSDb and a number of additional resources. We create our annotated corpora as follows:

1. First we collect **speaker-utterance pairs** from IMSDb scripts, based on their respective indentation sizes. We then use the Champollion sentence aligner (Ma, 2006) monolingually to align English subtitles to the English script, and follow the OpenSubtitles alignment links to align foreign subtitles to the English subtitles-script bitext. Finally, we discard the script text from the resulting 'tritext', yielding multilingual speaker-annotated dialogue corpora. The complete process of speaker-utterance annotation is graphically displayed in Figure 7.1. In addition, Table 7.1a shows statistics on the average number of speakers and main characters, i.e., speakers with at least 20 utterances, per movie.

2. We learn each speaker's **gender** based on their name's occurrence in a number of online name databases and a list of gender-revealing nouns such as 'aunt', 'boy',

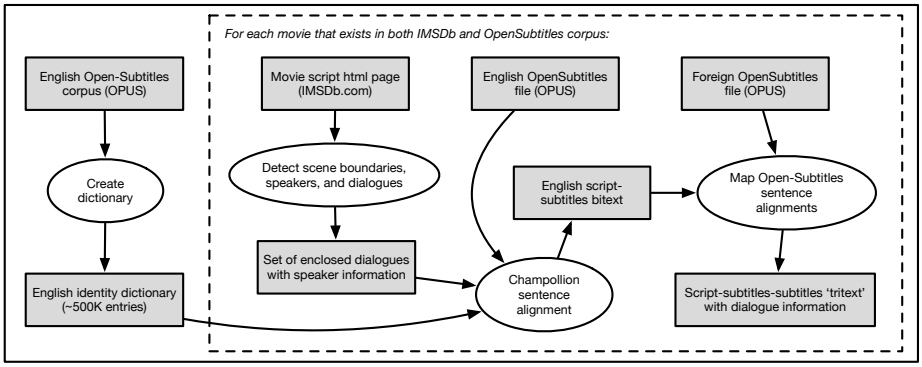


Figure 7.1: Diagram of the alignment process (step 1 in the complete annotation process) of English movie scripts, English subtitles, and foreign subtitles.

or 'grandma'. Annotations are available for ∼58% of the utterances, and the average female-to-male ratio is 1:1.7, see Table 7.1b.

3. We heuristically detect the **dialogue act** of each source-language utterance, considering *questions*, *exclamations* and *declaratives*. Distributions of these dialogue acts differ between language pairs, see Table 7.1c, but make up on average 28%, 9%, and 63% of the corpora, respectively.

4. We define a **register** label, based on the fraction of colloquial and vulgar expressions in an utterance, according to meta-information from an online dictionary.[3] We consider three register levels: *vulgar*, *colloquial*, and *neutral*, comprising circa 10%, 68%, and 22% of the corpora, respectively. Distribution statistics per language pair are shown in Table 7.1d.

7.3.2 Post-processing and annotation quality

The above described alignment and annotation process is done fully automatically, making it prone to errors. We therefore increase the alignment and annotation quality of our corpus by taking a number of measures: first, *before* running the Champollion aligner, we remove context information such as '[moaning]', '[clapping]' or '[chuckles]', which is most prevalent in, but not limited to, subtitles created for hearing-impaired people. In addition, we remove subtitle-specific tokens indicating continuation of a sentence on the next screen or switches in speaker turns, yielding more fluent and less fragmented utterances.

Next, *after* running the alignment process, we favor high-quality alignments by selecting only movies or movie versions (OpenSubtitles typically contains several alternative versions for a single movie (Tiedemann, 2016)) that meet the following criteria: (i) sentence lengths between both language pairs are sufficiently close, (ii) the number of sentences for which ambiguous speakers have been aligned does not

[3] `www.dict.cc`

Table 7.1: Annotation distributions of the dialogue benchmarks. a) Main characters are speakers with 20 or more utterances. b) Uncertain gender annotations are labeled 'unknown'. c) Annotated dialogue acts: questions, exclamations, declaratives. d) Annotated register levels: vulgar, colloquial, neutral.

	a) Avg. speaker statistics		b) Gender statistics			c) Dialogue act statistics			d) Register statistics		
Lang. pair	Speakers	Main chars	%M	%F	%Unk.	%Ques.	%Excl.	%Decl.	%Vulg.	%Coll.	%Neut.
AR → EN	44.6	5.6	36.2	20.4	43.4	29.8	6.0	64.2	10.8	67.1	22.1
DE ↔ EN	47.3	5.9	36.1	19.8	44.1	27.2	12.2	60.6	10.3	67.7	22.0
ES ↔ EN	42.5	6.0	37.2	22.9	39.9	28.4	11.2	60.4	10.6	67.7	21.7
NL ↔ EN	43.8	6.0	36.5	22.3	41.2	29.0	4.2	66.8	10.1	68.3	21.6
ZH → EN	42.3	5.6	36.9	21.1	42.0	28.1	10.6	61.3	10.7	68.9	20.4

Table 7.2: Confusion matrices for manual and automatic annotation of speaker gender (left; M=male, F=female, U=unknown), dialogue acts (center; Q=questions, E=exclamations, D=declaratives), and register levels (right; V=vulgar, C=colloquial, N=neutral). Overall percentage of agreement for are 86.7%, 97.5% and 85% for gender, dialogue act and register annotations, respectively.

Gender Annot.		Automatic M	F	U	Total
Manual	**M**	42	0	8	50
	F	1	27	3	31
	U	2	2	35	39
Total		45	29	46	120

Dial.act Annot.		Automatic Q	E	D	Total
Manual	**Q**	28	0	0	28
	E	1	8	0	9
	D	1	1	81	83
Total		30	9	81	120

Register Annot.		Automatic V	C	N	Total
Manual	**V**	9	0	0	9
	C	0	74	6	80
	N	0	12	19	31
Total		9	86	25	120

exceed a given threshold, (iii) the letter distribution is sufficiently similar to the average distribution of the language. By enforcing these quality standards, we respectively reduce the number of OpenSubtitles alignment errors, Champollion alignment errors, and OCR errors. Finally, we remove utterances with ambiguous speaker labels as these are caused by erroneous Champollion alignments.

Despite efforts to improve the alignment quality, our corpus still contains some incorrect alignments due to either OpenSubtitles or Champollion errors. To quantify these, and to verify the correctness of the automatic annotations, we manually inspect randomly selected fragments across different language pairs and movies. Based on evaluation of a sample of 120 utterances, we estimate a final alignment accuracy of 92.5%. In addition, Table 7.2 shows confusion matrices for manual versus automatic annotation of speaker gender, dialogue acts, and register for the 120 selected utterances. With the overall annotation agreement per variable ranging from 85% to 97.5%, we find that our automatic annotation strategies are very accurate. Disagreement between manual and automatic annotations occurs mostly for speakers labeled with unknown gender, and between the register levels colloquial and neutral, indicating that these categories can benefit from more advanced annotation methods.

Table 7.3 shows an example dialogue with annotations and its original form in the OpenSubtitles corpus. Note that our annotated corpora differ from OpenSubtitles since only actual dialogues are included, many erroneously aligned sentence pairs have been removed, and utterances are longer and less fragmented. The latter is a result of the Champollion alignment process. Since sentences in the IMSDb scripts are typically longer than those in OpenSubtitles, Champollion regularly enforces one-to-many alignments. Following the OpenSubtitles-internal alignment links then often yields many-to-many alignments in which subtitles get merged into longer utterances.

Finally, Table 7.4a lists the statistics of the benchmarks which we use in this chapter and make available for download.[4] While the remainder of this chapter uses the annotated fictional dialogues to analyze the impact of dialogue-specific aspects on PBMT, we believe that our data set may also help to advance dialogue research—today largely confined to the English language—in a multilingual scenario.

7.4 Measuring dialogue effects on PBMT

In this section we measure the effect of dialogue dimensions on PBMT performance of fictional dialogues. To this end, we quantify BLEU fluctuations between differences in dialogue acts, speakers, gender, and register, and we determine whether the observed fluctuations are larger than randomly expected.

7.4.1 Basic experimental setup

We run our experiments using Oister (see Section 2.4.1). Our systems are trained on 59M–171M tokens (depending on the language pair, see Table 7.4b) of unannotated OpenSubtitles corpora. We use 5-gram language models trained on 500M–1.7B tokens, depending on the language pair, that linearly interpolate OpenSubtitles with various

[4] http://ilps.science.uva.nl/resources/movie-dialogues

Table 7.3: Example dialogue from Notting Hill: a) original sentences in the OpenSubtitles corpus, where // indicates a sentence boundary in many-to-one or one-to-many alignments, and b) final annotated utterances generated in our annotation pipeline, annotated with speaker (William, Martin), gender (M=male, F=female), register level (neutral, colloquial, vulgar) and dialogue act (declarative, exclamation, question). Note that sentences pairs from the original corpus are often merged in the Champollion alignment process, and that erroneous OpenSubtitles alignments are not corrected.

a) Original German-English OpenSubtitles alignment

German subtitles	**English subtitles**
Erstklassig!	Classic.
Bilanz der Werbekampagne... minus 347 Pfund.	Profit from major sales push, minus £347.
Soll ich... dir einen Cappuccino holen?	Shall I go and get you a cappuccino? // You know, ease the pain a bit. // Yeah.
Als Seelentröster?	
Ja.	Yeah.
Lieber nur einen halben.	Better make it a half.
Mehr kann ich mir nicht leisten.	All I can afford.
Logisch.	Get your logic.
Demi-Cappu. // Kommt sofort.	Demi-cappu coming right up.

b) Annotated German-English dialogue

German utterance	**English utterance**	**Annotations**
Erstklassig! Bilanz der Werbekampagne minus 347 Pfund.	Classic. Profit from major sales push, minus £347.	William, M, neutral, exclamation
Soll ich dir einen Cappuccino holen?	Shall I go and get you a cappuccino? You know, ease the pain a bit.	Martin, M, coll., question
Ja. Lieber nur einen halben. Mehr kann ich mir nicht leisten.	Yeah. Better make it a half. All I can afford.	William, M, coll., declarative
Logisch. Demi-Cappu. Kommt sofort.	Get your logic. Demi-cappu coming right up.	Martin, M, coll., declarative

Table 7.4: Specifications of parallel training and evaluation data. Training data consists of OpenSubtitles corpora, evaluation data consists of speaker-annotated dialogues.

	a) Evaluation data		b) Training data	
Languages	Movies	Utterances	Lines	EN tokens
AR → EN	187	94K	12.3M	122M
DE ↔ EN	220	123K	9.7M	85M
ES ↔ EN	161	87K	16.3M	156M
NL ↔ EN	238	129K	17.9M	171M
ZH → EN	211	107K	6.3M	59M

LDC and WMT corpora using weights optimized on a held-out set of OpenSubtitles data. Systems are tuned on a different held-out OpenSubtitles set. The resulting systems are thus at all levels adapted to the movie dialogues translation task rather than the general domain.

7.4.2 Approximate randomization testing

When translating dialogues, we naturally observe *some* BLEU variations across categories such as different dialogue acts or speakers. An important question is whether the observed differences are to be expected (the null hypothesis), or whether they are indicators that one category is truly harder to translate than another (the alternative hypothesis). We test this hypothesis with an approximate randomization approach (Edgington, 1969; Noreen, 1989).

While approximate randomization (also known as approximate permutation) is often used to compare the mean and variance of *two* groups, it can be adapted to our setting with multiple categories. To this end, we compute BLEU for each of the categories in a dialogue variable, e.g., vulgar, colloquial, and neutral utterances for the dialogue variable of register level. Next, we randomly permute category labels over utterances, following the original distribution of utterances per category, and we recompute BLEU for the randomized labels.

As our test statistic of interest, we define and measure the *mean absolute BLEU difference*, which captures BLEU fluctuations between categories:

$$\text{MBD} = \frac{2}{|S|^2 - |S|} \sum_{i=1}^{|S|} \sum_{j=i+1}^{|S|} \mid \text{BLEU}_i - \text{BLEU}_j \mid \tag{7.1}$$

Here BLEU_i the BLEU score for category i, and S is the set of categories for a given dialogue variable, e.g., $S_{\text{register}} = \{\text{vulgar}, \text{colloquial}, \text{neutral}\}$. Each pair of categories (i, j) is compared exactly once in terms of BLEU scores. The sum of all pairwise BLEU differences is normalized for the total number of unordered pairs (i, j), which is $\binom{|S|}{2} = \frac{|S|^2 - |S|}{2}$. Note that MBD is a specific instance of *mean absolute difference* (MD), also known as *Gini mean absolute difference* (GMD), a measure of statistical

dispersion which has shown to be superior to other common statistical dispersion measures such as variance, standard deviation and interquartile range (Yitzhaki, 2003).

Next, we compute the p-value by counting how often, in a total of 1,000 permutations, we observe an MBD value that is at least as extreme as the one observed for the real categories. If for a given dialogue variable $p \leq 0.05$ or $p \leq 0.01$, we conclude that this variable has a weakly or strongly significant impact on PBMT quality, respectively.

For dialogue acts, gender and register we permute labels over the entire benchmark. For speakers we only permute labels within each movie since inter-movie variations in BLEU are affected by many other factors, e.g., script writers, translators, movie genre, which distract from the impact of speakers. In addition, when computing speaker-specific BLEU, we only include main characters, i.e., speakers with at least 20 utterances, to avoid BLEU's instability on small documents.

7.5 Results

In this section we discuss the observed BLEU fluctuations (see Table 7.5) for our four dialogue variables of interest. We provide pointers in the text to the phenomena observed in the Spanish-English and English-German examples in Table 7.6.

7.5.1 The effect of dialogue acts on PBMT quality

As shown in Table 7.5a, there are substantial performance fluctuations between dialogue acts for all language pairs. However, there is no consistent pattern between different languages. For instance, we observe punctuation errors (EX1) for Spanish-English and English-Spanish translation, and verb drop (EX2) and wrong word order (EX3) for English-German translation.

Since there is no consistent pattern among language pairs, we want to verify whether BLEU fluctuations can be explained by variations in average sentence length. To this end, we compute the Pearson correlation between the average sentence length for each dialogue act and the corresponding BLEU scores, all normalized per language pair. We find a weak negative correlation of $r = -0.11$, indicating that shorter sentences tend to have slightly higher BLEU scores, but also that sentence length variations alone can not explain the large BLEU variations. This makes it particularly interesting to further investigate how dialogue acts can be exploited to improve translation quality of (fictional) dialogues. However, when considering finer dialogue act granularities, it may be profitable to exploit context information, i.e., previous and following sentences, which is not used in our current PBMT setup.

Finally, improving PBMT for the dialogue acts under consideration can also be useful for cross-lingual question answering (Tiedemann, 2009b; Ture and Boschee, 2016).

7.5.2 The effect of speakers on PBMT quality

In Table 7.5b we report the percentage of movies per language pair for which the observed MBD is statistically significant at $p \leq 0.05$, which is 18.6% on average. Since

Table 7.5: BLEU for dialogue acts, speakers, gender, and register, translated using baseline PBMT trained and tuned on OpenSubtitles corpora. MBD: mean absolute BLEU difference, see Equation (7.1), all statistically significant at $p \leq 0.01$ (▲). ssMBD: percentage of movies with statistically significant speaker-MBD at $p \leq 0.05$.

	a) BLEU per dialogue act				b) Speaker-ssMBD	c) BLEU per gender			d) BLEU per register			
Lang.pair	Quest.	Excl.	Decl.	MBD		Male	Female	MBD	Vulg.	Coll.	Neut.	MBD
AR→EN	23.1	20.3	19.3	2.5▲	14.9%	20.4	22.0	1.6▲	17.2	21.3	19.7	2.8▲
DE→EN	24.0	20.9	21.4	2.0▲	22.3%	21.4	22.7	1.2▲	17.7	21.9	24.7	4.6▲
ES →EN	28.9	25.7	26.8	2.1▲	18.0%	26.7	28.0	1.3▲	24.4	27.4	28.8	2.9▲
NL→EN	26.7	29.0	23.7	3.5▲	23.9%	23.9	26.8	2.9▲	20.8	24.8	26.7	4.0▲
ZH→EN	15.3	15.2	12.9	1.6▲	16.6%	12.8	14.4	1.6▲	10.9	13.4	13.1	1.7▲
EN→DE	17.7	18.5	16.5	1.3▲	15.9%	16.7	17.6	0.9▲	13.6	16.6	19.9	4.2▲
EN→ES	16.8	16.5	21.5	3.3▲	13.7%	18.9	19.7	0.8▲	17.2	19.2	21.0	2.6▲
EN→NL	25.5	22.7	24.6	1.8▲	20.6%	23.9	26.5	2.6▲	21.4	24.6	26.3	3.3▲

Table 7.6: Censored ES→EN (top) and EN→DE (bottom) translation examples of an annotated dialogue, originating from Pulp Fiction and involving two speakers: Pumpkin (S1, M=male) and Honeybunny (S2, F=female). Examples of phenomena marked with (EX#) are discussed in Sections 7.5.1–7.5.4.

Annotations	ES source	ES→EN PBMT output	EN reference
S1, M, neutral, declarative	se acabaron los días de olvidar, han empezado los de recordar.	the days are over, have begun to forget them to remember.	the days of forgetting are over. the days of remembering have begun.
S2, F, colloquial, question	¿sabes qué pareces cuando hablas así?	you know what you look like when you talk like that?	know when you go on what you sound like?
S1, M, vulgar, declarative	un *j***do*(EX4) hombre sensato.	a sensible man.(EX6)	i sound like a sensible *f***ing* man.
S2, F, colloquial, excl.	un pato. ¡cuac, cuac!	a duck. [quack, quack!(EX1)	you sound like a duck. quack, quack

Annotations	EN source	EN→DE PBMT output	DE reference
S1, M, neutral, declarative	the days of forgettin' are over. the days of remembering have begun.	die tage von vergisst sind vorbei. die tage von an(EX2) haben begonnen.	ja, aber jetzt kommen die tage des erinnerns.
S2, F, colloquial, question	know when you go on what you sound like?	weiß, wenn du auf(EX2) was du klingst wie?(EX3)	weißt du, wie du klingst?
S1, M, vulgar, declarative	i sound like a sensible *f***ing* man.	ich klinge wie ein vernünftig *verd***ter* mann.	wie ein vernünftiger(EX5) mann.
S2, F, colloquial, declarative	you sound like a duck. quack, quack	du klingst wie eine ente. quak, quak	nein, wie eine ente! quak, quak, quak!

there are too many speakers to report individual BLEU scores, we randomly select 100 German-English movies, and compute for each of these ΔMBD as the difference between MBD for real speakers and the average MBD for randomized labels:

$$\Delta\text{MBD} = \text{MBD}_{real} - \overline{\text{MBD}}_{random} \tag{7.2}$$

Figure 7.2 shows that BLEU fluctuations among real speakers are often larger than BLEU fluctuations among randomized speaker tags. These findings suggest that, while domain adaptation is an established task in PBMT, conversational PBMT may benefit—at least for the fraction of movies with statistically significant speaker differences—from a fine-grained adaptation at the speaker level, as proven successful in speech recognition research (Shinoda, 2011), and related to recent work on personalizing machine translation (Mirkin et al., 2015; Mirkin and Meunier, 2015). For movies with negative ΔMBD, BLEU variation between different speakers is lower than randomly expected, hence successful adaptation should take place at levels other than speakers.

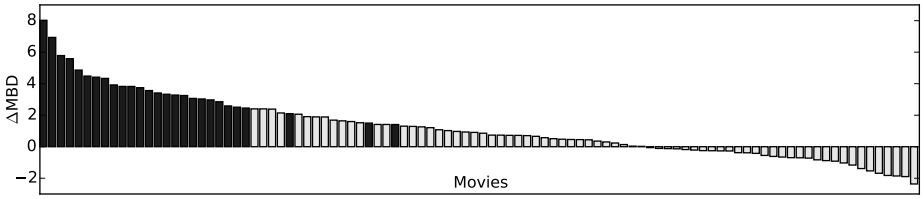


Figure 7.2: ΔMBD (see Section 7.5.2) for 100 randomly selected German-English benchmark movies. Black bars indicate movies with statistically significant positive ΔMBD at $p \leq 0.05$.

Finally, since our analysis is carried out on fictional dialogues, it may be worth investigating to what extent BLEU scores fluctuate between actors or script writers rather than only characters, however this requires additional annotation.

7.5.3 The effect of gender on PBMT quality

Table 7.5c shows that BLEU scores per gender follow a similar pattern in all language pairs. Male speakers are significantly harder to translate than female speakers, despite the fact that male speakers are likely better represented in the parallel OpenSubtitles data, based on the male-to-female ratio of 1.7:1 in our evaluation sets. However, we find that female utterances are better covered by the language model, with perplexity values on average 8% higher for males than females. In addition, female utterances are on average slightly shorter than male utterances, i.e., 11.8 versus 12.3 tokens for female and male utterances, respectively, and may therefore be slightly easier to translate. Interestingly, Bawden et al. (2016) find that PBMT adaptation towards speaker genders mostly benefits from an increased average length of output translations, however in their setting female utterances are harder to translate than male utterances.

Gender differences in fictional dialogues have also been reported by Danescu-Niculescu-Mizil and Lee (2011), who show that movie characters mirror their language use more easily to females than to males. However, we have to be careful when drawing conclusions about the impact of gender on real spoken dialogues from our

observations on fictional dialogues. Since the vast majority of movie scripts are written by men (Lauzen, 2016), our findings reflect differences between language of male and female characters as perceived by male writers. The observed BLEU differences might therefore be based on stereotypes rather than real gender differences. On the other hand, a large body of work has studied gender and language or discourse (Tannen, 1994; Wodak, 1997; Holmes and Meyerhoff, 2008, among others), indicating that these concepts are closely intertwined. We therefore believe that meta-information about a speaker's gender is also a potential source to customize PBMT for real dialogues.

7.5.4 The effect of register on PBMT quality

The results per register (Table 7.5d) show that PBMT quality is worst for vulgar utterances and generally best for neutral sentences. While consistent with our previous findings that informal language is hard to translate (see Chapters 3 and 6), this observation cannot solely be attributed to poor model coverage, since colloquial and vulgar language are well-represented in our OpenSubtitles-trained systems. However, due to our annotation procedure, short sentence that are easy to translate are mostly marked as neutral, hence increasing BLEU for neutral utterances. This indicates that register annotations may be improved for example using LM rather than dictionary lookup techniques.

When manually inspecting human translations for vulgar expressions, we find that these vary from literal (EX4) to very nuanced (EX5) translations, yielding inconsistent PBMT output. We also observe that vulgarity is often not preserved in (both human and machine) translation (EX6): A comparison of vulgarity scores shows that, while vulgarity sometimes increases, the number of vulgar utterances in the PBMT output is on average 35% lower than in the reference set.

Finally, since poor PBMT quality is observed for both male characters and vulgar language, we hypothesize that the two might co-occur. Indeed, the average vulgarity scores for males are 64% higher than for females, which may in part explain the observed PBMT quality between genders.

7.6 Adaptation towards dialogue variables

We observed in the previous section that BLEU scores significantly fluctuate along dialogue dimensions. This finding suggests that PBMT for fictional dialogues may benefit from adaptation towards different categories along these dimensions. To verify this hypothesis we run a number of adaptation experiments, in which we adapt our baseline PBMT systems towards different dialogue acts and different registers—two dialogue aspects which can be computed straightforwardly for the unannotated training corpora.

We adapt our systems at two levels: first, we create category-specific language models by interpolating our general movie dialogue language model with a language model trained on only the most relevant subset of the bitext's target side. We determine relevant sentences by applying the same annotation guidelines that were used for

Table 7.7: Results of adaptation experiments. Top: adaptation towards dialogue acts for the 3 language pairs with the largest mean absolute BLEU difference (MBD, see Equation (7.1)) between dialogue acts. Bottom: adaptation towards registers for the 3 language pairs with the largest MBD between register levels. Statistical significance against the baseline at $p \leq 0.05$ ($\triangle$/$\triangledown$) and $p \leq 0.01$ (▲/▼) is measured using approximate randomization (Riezler and Maxwell, 2005).

	MBD		Questions			Exclamations			Declaratives		
Language pair	Base.	Adapt.	Base.	Adapt.	Diff.	Base.	Adapt.	Diff.	Base.	Adapt.	Diff.
NL → EN	3.5	3.2	26.7	26.5	–0.2$^{\triangledown}$	29.0	28.9	–0.1	23.7	24.1	+0.4$^{\blacktriangle}$
EN → ES	3.3	2.8	16.8	17.5	+0.7$^{\blacktriangle}$	16.5	17.3	+0.8$^{\blacktriangle}$	21.5	21.5	0.0
AR → EN	2.5	2.6	23.1	24.3	+1.2$^{\blacktriangle}$	20.3	20.4	+0.1	19.3	20.9	+1.6$^{\blacktriangle}$

	MBD		Vulgar			Colloquial			Neutral		
Language pair	Base.	Adapt.	Base.	Adapt.	Diff.	Base.	Adapt.	Diff.	Base.	Adapt.	Diff.
DE → EN	4.6	4.2	17.7	18.4	+0.7$^{\blacktriangle}$	21.9	22.7	+0.8$^{\blacktriangle}$	24.7	24.7	0.0
EN → DE	4.2	3.8	13.6	14.6	+1.0$^{\blacktriangle}$	16.6	17.8	+1.2$^{\blacktriangle}$	19.9	20.3	+0.4$^{\blacktriangle}$
NL → EN	4.0	2.8	20.8	23.1	+2.3$^{\blacktriangle}$	24.8	25.6	+0.8$^{\blacktriangle}$	26.7	27.3	+0.6$^{\blacktriangle}$

annotation of the benchmarks (Section 7.3). Second, we tune our systems on held-out sets selected according to the same criteria, thus comprising category-specific data.

We run adaptation experiments for the language pairs with the largest observed MBD: Dutch-English, English-Spanish, and Arabic-English for dialogue acts, and German-English, English-German, and Dutch-English for register. Note that the aim of our adaptation experiments is to verify whether PBMT performance can benefit from a simple adaptation approach at the fine-grained level of different dialogue-specific aspects, rather than presenting a novel PBMT adaptation approach.

The results of our adaptation experiments are shown in Table 7.7. The first observation we can make is that the adapted systems result in substantially lower mean absolute BLEU differences (MBD) for both dialogue dimensions—dialogue act and register level—for all language pairs except Arabic-English. This means that most of the adapted systems generate translations of more uniform quality with a lower degree of fluctuation in BLEU. Further, the BLEU scores for the individual categories of both dialogue dimensions show that the lower MBD scores are due to statistically significant improvements for most of the dialogue acts and registers. The only case where our simple adaptation method causes a statistically significant drop in BLEU is for the translation of questions from Dutch into English. Vulgar and colloquial language profit particularly well from language model adaptation, while results for question and exclamation marks are more variable between language pairs.

7.7 Conclusions

In previous PBMT research, very little work has been reported on PBMT for informal genres such as dialogues, a genre that differs substantially from more formal text and thus poses different translation challenges. Dialogues involve, by definition, multiple *speakers*, often with different *genders*, who likely have varying intentions and language use. Dialogues are therefore characterized by functional actions known as *dialogue acts* and utterances with varying *register* levels. While these and other dialogue-specific aspects have been analyzed extensively in dialogue research (Fernández, 2014), their impact on PBMT has hardly been studied. In this chapter we addressed this issue and investigated the effect of four dialogue-specific aspects on PBMT quality, by asking:

RQ5 *What impact do dialogue-specific aspects have on translation quality?*

As a first step towards answering this question, we created movie-dialogue benchmarks for five language pairs in which utterances are annotated with dialogue acts, speakers, speaker gender, and register. Armed with this new resource, we found that BLEU fluctuations for speaker gender, dialogue acts, and register levels are always significantly larger than randomly expected. As for speaker variations, we found that on average 18.6% of the movies show larger fluctuations than expected at random, suggesting that for this subset of movies adaptation towards speakers may benefit PBMT quality.

When examining the impact of specific dialogue aspects, we observed that the register level has a significant impact on translation quality, with translations of vulgar utterances being of substantially lower quality than neutral or even colloquial utterances. Similarly we found large variations in translation quality between different dialogue

acts, although we did not detect a consistent pattern between different languages: e.g., questions can be more difficult to translate than exclamations for one language pair, while the reverse is true for another language pair. Finally, we found that male speakers are harder to translate and use more vulgar language than female speakers, and that vulgarity is often not preserved during translation.

Our observations suggest that conversational PBMT may benefit from adaptation at fine-grained levels, which we tested and confirmed in a series of simple adaptation experiments towards different dialogue acts and different register levels. This indicates that apart from domain adaptation, adaptation to other variables should be considered to improve PBMT quality for fictional dialogues. While our analyses are carried out on fictional dialogues, we believe that our findings generalize at least partially to other types of dialogues, and are thus valuable for advancing conversational PBMT.

Part III

Domain Adaptation for Neural Machine Translation

One of the goals of this thesis is to develop and present methods for domain or genre adaptation that do not—or at most partially—depend on the availability of subcorpus information. The reasons for this are twofold: first, subcorpus information may not be available, for example in online scenarios or when training data is automatically harvested from the web. Second, subcorpus information may not provide the most informative categorization. For example, we saw in Chapter 4 that 'misclassification' of a genre label can in fact improve PBMT quality, and in Chapter 5 that a subcorpus or genre grouping that is informative for one translation task does not always generalize to other tasks.

After the adaptation methods using automatic genre classifiers (Chapter 4) or textual features (Chapter 5) presented in part I, we target in the last part of this thesis domain adaptation by means of automatic training data selection. Specifically, given a translation task, we automatically select the most relevant training data using a language model-based selection strategy. With this method we do not need any information on the origin of the training data. In addition, we only need small amounts of data representing the 'domain' of the test set.

Data selection has proven a successful technique to simultaneously increase training efficiency and translation performance for domain adaptation in PBMT. With the recent increase in popularity of NMT, we explore in this part *to what extent* and *how* NMT can also benefit from data selection. Since domain adaptation for NMT is still an under-explored research direction, we evaluate our methods with domains that are defined by their provenance, and we leave fine-grained levels of adaptation for future work.

8

Dynamic Data Selection for Neural Machine Translation

8.1 Introduction and research questions

Recent years have shown a rapid shift from PBMT to NMT (Sutskever et al., 2014; Cho et al., 2014b; Bahdanau et al., 2014) as the most common machine translation paradigm. With large quantities of parallel data, NMT outperforms PBMT for an increasing number of language pairs (Bojar et al., 2016). Unfortunately, training an NMT model is often a time-consuming task, with training times of several weeks not being unusual.

Despite its training efficiency, most work in NMT greedily uses all the available training data for a given language pair, regardless of the translation task of interest. However, it is unlikely that all data is equally helpful to create the best-performing system for a given domain. In PBMT, this issue has been addressed by applying *data selection*, and it has consistently been shown that using more data does not always improve translation quality (Moore and Lewis, 2010; Axelrod et al., 2011; Gascó et al., 2012). Instead, for a given translation task, the training bitext likely contains sentences that are irrelevant or even harmful, making it beneficial to discard these sentences.

While data selection can be performed using domain information of the training and test data, it can also be done in an automated fashion, where one does not need to know the provenance of training data in advance. In general, automatic data selection is preferable over manual data selection since (i) it allows for the inclusion of training data with unknown origin, and (ii) the most relevant training data may not exactly be the subset labeled as 'in-domain'. We have seen an example of the latter in Chapter 4 when misclassification of a document's genre in fact led to improved BLEU scores.

Motivated by the success of data selection in PBMT, we investigate in this chapter whether NMT can benefit from data selection as well, in particular in the context of domain adaptation. To this end, we ask:

RQ6 *To what extent and how can we successfully apply data selection for domain adaptation in NMT?*

While data selection has been applied to NMT to reduce the size of the data (Cho et al., 2014b; Luong et al., 2015b), the effects on translation quality have not been investigated.

Intuitively, data selection for NMT is a challenging task: NMT systems are known to under-perform when trained on limited parallel data (Zoph et al., 2016; Fadaee et al., 2017; Koehn and Knowles, 2017), and do not have a separate large-scale target-side language model to compensate for smaller parallel training data. To gain insights into the effects of data selection on NMT, we first apply a state-of-the-art data selection method (Axelrod et al., 2011) to both PBMT and NMT, and we ask:

RQ6a. *How does a state-of-the-art data selection approach perform when applied to PBMT and NMT?*

Next, to alleviate the negative effect of small training data on NMT, we introduce *dynamic data selection*, and we ask:

RQ6b. *Can we improve upon state-of-the-art data selection for NMT by dynamically changing the selected data subsets during training?*

Following conventional data selection, we still dramatically reduce the training data size, favoring parts of the data which are most relevant to the translation task at hand. However, we exploit the fact that the NMT training process iterates over the training corpus in multiple epochs, and we alter the quantity or the composition of the training data *between epochs*. We propose two variants of dynamic data selection, *sampling* and *gradual fine-tuning*. Dynamic data selection requires no modifications to the NMT architecture or parameters, and substantially speeds up training times while simultaneously improving translation quality with respect to a complete-bitext baseline.

Organization This chapter is organized as follows: in Section 8.2 we discuss related work. In Section 8.3, we describe the state-of-the-art data selection method that we use for comparison between PBMT and NMT. Next, in Section 8.4, we introduce dynamic data selection as a way to make data selection more effective for NMT. We describe our basic data and experimental settings in Section 8.5, and discuss the results of our experiments in Section 8.6. In Section 8.7, we further analyze the results obtained with dynamic data selection. Finally, we provide conclusions in Section 8.8.

8.2 Related work

A number of research topics are related to the work in this chapter. First, we provide an overview of work on data selection for PBMT, which has has been extensively studied in the literature. Next, since data selection can simultaneously serve two goals, improving domain adaptation and increasing training efficiency, we discuss either goal and its related work to date in NMT.

8.2.1 Data selection for PBMT

Regarding data selection for PBMT, previous work has targeted the two goals mentioned above: reducing model sizes and training times, for example by speeding up model building and reducing the complexity of search through the decoding space, or adapting

to new domains. Data selection methods for the purpose of domain adaptation mostly employ information theory metrics to rank training sentences by their relevance to the domain at hand. This has been applied monolingually (Gao et al., 2002) as well as bilingually (Yasuda et al., 2008). In more recent work, training sentences are typically ranked according to their cross-entropy *difference* between in-domain and general-domain data (Moore and Lewis, 2010; Axelrod et al., 2011, 2015), favoring sentences that are similar to the test domain and at the same time dissimilar from the general domain. Using such a method, Axelrod et al. (2011) have shown that using only 1% of a very large and diverse training corpus can yield higher BLEU than using the entire bitext. Duh et al. (2013) and Chen and Huang (2016) present similar methods in which n-gram LMs are replaced by recurrent neural LMs or neural classifiers, respectively.

Data selection with the aim of model size and training time reduction has the objective to use a minimum amount of data while still maintaining high vocabulary coverage. This strategy has been applied by Eck et al. (2005) and Gascó et al. (2012), who rank sentences based on the frequencies of unseen n-grams, and by Lewis and Eetemadi (2013), who set a limit on the number of times each n-gram can be selected. In a comparative study, Mirkin and Besacier (2014) find that data selection with the objective to select data that is similar to the test set performs best if the test domain and general corpus are very different, while data selection with the objective to maintain high coverage is superior if the test set and the general corpus are relatively similar. A comprehensive survey on data selection for PBMT is provided by Eetemadi et al. (2015).

While in this chapter we use a similarity objective to rank our bitext, one could also apply dynamic data selection using a coverage objective.

8.2.2 Domain adaptation for NMT

In NMT, like in PBMT, one of the goals of data selection could be to apply domain adaptation. To date, domain adaptation in NMT typically involves training a model on the complete bitext, followed by fine-tuning the parameters on a smaller in-domain corpus (Luong and Manning, 2015; Zoph et al., 2016). Freitag and Al-Onaizan (2016) combine fine-tuning with model ensembles, and Chu et al. (2017) with domain-specific tags in the training corpus, which is based on a similar method that uses domain labels without fine-tuning (Kobus et al., 2016). Sennrich et al. (2016b) adapt their systems by back-translating monolingual in-domain data, which is then added to the training data and also used for fine-tuning. The disadvantage of the above mentioned domain-specific fine-tuning approaches is that translation quality severely degrades for other domains. Dakwale and Monz (2017) address this issue by using supervision from a general-domain 'teacher' network during fine-tuning, ensuring limited deterioration of NMT quality on out-of-domain test data.

A number of very recent approaches have targeted domain adaptation for NMT in a different way than using fine-tuning. For example, Kreutzer et al. (2017) apply bandit learning with weak feedback to learn in-domain model parameters. Hokamp and Liu (2017) introduce 'Grid Beam Search,' an extension of beam search that allows one to lexically constrain decoding to pre-defined lexical, e.g., in-domain, constraints. Farajian et al. (2017) show that multi-domain NMT, in which a single system is tailored

to translate a variety of domains, is still far behind compared to multi-domain PBMT.

Finally, Wang et al. (2017a) also apply data selection for NMT. Their approach is inspired by Axelrod et al. (2011), however rather than cross-entropy differences, they use the NMT-internal embedding of each source sentence to score the corresponding sentence pair's similarity to the domain of interest. In related work, Wang et al. (2017b) apply instance weighting to domain adaptation for NMT, by integrating relevance weights (based on sentence-level cross-entropy scores or based on domain labels) into the NMT objective function.

8.2.3 Training efficiency for NMT

The other benefit of applying data selection is the reduction of training times, which can be extremely large for NMT—training times up to several weeks have been reported. To this end, a number of previous efforts have addressed training efficiency for NMT, for example by parallelizing models or data (Wu et al., 2016), modifying the NMT network structure (Kalchbrenner et al., 2016), decreasing the number of parameters through knowledge distillation (Crego et al., 2016; Kim and Rush, 2016), or by boosting parts of the data that are 'challenging' to the NMT system (Zhang et al., 2016). The latter is closely related to our work since training data is also adjusted during training, however we reduce the training data size much more aggressively and study different techniques of data selection.

While data selection for PBMT has the additional advantage of reduced model sizes, the size of an NMT model is defined by the model architecture and parameters, and is thus not affected by data selection.

8.3 Static data selection

As a first step towards dynamic data selection for NMT, we compare the effects of a commonly used, state-of-the-art data selection method (Axelrod et al., 2011) on both neural and phrase-based MT. Briefly, this approach ranks sentence pairs in a large training bitext according to their difference in cross-entropy with respect to an in-domain corpus, i.e., a corpus representing the test data, and a general corpus. Next, the top n sentence pairs with the highest rank—thus lowest cross-entropy—are selected and used for training an MT system. Following common conventions, we refer to data resembling the test data as *in-domain* data.

Formally, given an in-domain corpus I, we first create language models from the source side f of I ($LM_{I,f}$) and the target side e of I ($LM_{I,e}$). We then draw a random sample (similar in size to I) of the large general corpus G and create language models from the source and target sides of G: $LM_{G,f}$ and $LM_{G,e}$, respectively. Note that the data for creating these LMs need not be parallel but can be independent corpora in both languages.

Next, we compute for each sentence pair s in G four cross-entropy scores:

$$H_{C,s_b} = -\sum p\left(s_b\right) \log\left(LM_{C,b}\left(s_b\right)\right), \tag{8.1}$$

where $C \in \{I, G\}$ is the corpus, $b \in \{f, e\}$ refers to the bitext side, and s_b is the bitext side b of sentence pair s in the parallel training corpus.

To find sentences that are similar to the in-domain corpus, i.e., have low H_I, and at the same time dissimilar to the general corpus, i.e., have high H_G, we compute for each sentence pair s the bilingual cross-entropy difference CED_s following Axelrod et al.:

$$\mathit{CED}_s = \left(H_{I,s_f} - H_{G,s_f}\right) + \left(H_{I,s_e} - H_{G,s_e}\right). \tag{8.2}$$

Finally, we rank all sentence pairs $s \in G$ according to their CED_s, and then select only the top n sentence pairs with the lowest CED_s.

Following related work by Moore and Lewis (2010), we restrict the vocabulary of the LMs to the words occurring at least twice in the in-domain corpus. Our LMs are 5-gram and created using the SRILM toolkit (Stolcke et al., 2002). To analyze the quality of the selected data subsets, we also run experiments on random selections, all performed in threefold to obtain stable results. Finally, we always use the exact same selection of sentence pairs in equivalent PBMT and NMT experiments.

LSTM versus n-gram The described data selection method uses n-gram LMs to determine the domain-relevance of sentence pairs. We adhere to this setting for our comparative experiments on PBMT and NMT (Section 8.6.1). However, when applying data selection to NMT, we examine the potential benefit of replacing the conventional n-gram LMs with LSTMs.[1] These have the advantage that they remember longer histories, and do not have to back off to shorter histories when encountering out-of-vocabulary words (Hochreiter and Schmidhuber, 1997; Mikolov, 2010). In this neural variant to rank sentences, the score for each sentence pair in G is still computed as the bilingual cross-entropy difference in Equation (8.2). In addition, we use the same in-domain and general corpora as with the n-gram method, and we again restrict the vocabulary to the most frequent words.

8.4 Dynamic data selection

While data selection aims to discard irrelevant data, it can also exacerbate the problem of low vocabulary coverage and unreliable statistics for rarer words in the 'long tail,' which are a major issue in NMT (Luong et al., 2015b; Sennrich et al., 2016c). In addition, it has been shown that NMT performance drops tremendously in low-resource scenarios (Zoph et al., 2016; Fadaee et al., 2017; Koehn and Knowles, 2017).

To overcome this problem, we introduce *dynamic data selection*, in which we vary the selected data subsets *during* training. Unlike other MT paradigms, which require training data to be fixed during the entire training process, NMT iterates over the training corpus in several epochs, allowing to use a different subset of the training data in every epoch.

Dynamic data selection starts from a relevance-ranked bitext, which we create using CED scores as computed in (8.2). Given this ranking, we investigate two dynamic data

[1] We use four-layer LSTMs with embedding and hidden sizes of 1,024, which we train for 30 epochs.

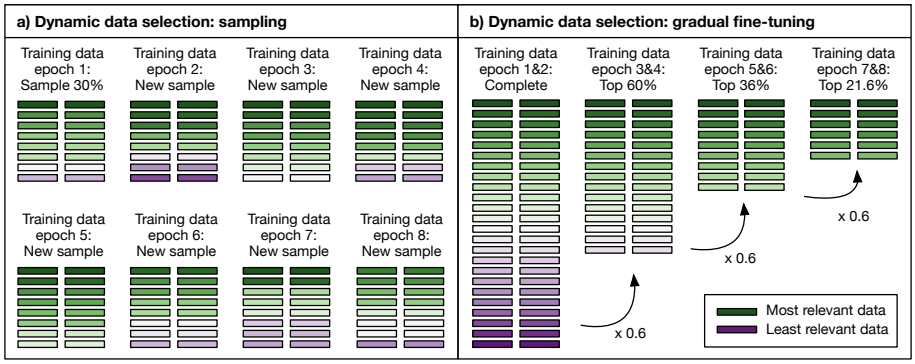


Figure 8.1: Illustration of two dynamic bitext selection techniques for NMT: *sampling* (left) and *gradual fine-tuning* (right). Measured over 16 epochs, as used in this work, the total training time of both examples would be ∼30% of the training time needed when using the complete bitext.

selection techniques[2] that vary per epoch the composition or the size of the selected training data. Both techniques aim to favor highly relevant sentences over less relevant sentences while not completely discarding the latter. In all experiments, we use a fixed vocabulary created from the complete bitext.

While in this work we use a domain-relevance ranking of the bitext following Axelrod et al., dynamic data selection can also be applied using other ranking criteria, for example limiting redundancy in the training data (Lewis and Eetemadi, 2013) or complementing similarity with diversity (Ruder and Plank, 2017).

8.4.1 Sampling sentence pairs

In the first technique, illustrated in Figure 8.1a, we sample for every epoch n sentence pairs from G, using a distribution computed from the domain-specific CED_s scores. Concretely, this is done as follows: first, since higher ranked sentence pairs have lower CED_s scores, and they can be either negative or positive, we scale and invert CED_s scores such that $0 \leq CED'_s \leq 1$ for each sentence pair $s \in G$:

$$CED'_s = 1 - \frac{CED_s - \min(CED_G)}{\max(CED_G) - \min(CED_G)}, \tag{8.3}$$

where CED_G refers to the set of CED_s scores for bitext G. Next, we convert CED'_s scores to relative weights, such that $\sum_{s \in G} w(s) = 1$:

$$w(s) = \frac{CED'_s}{\sum_{s_i \in G} CED'_{s_i}}. \tag{8.4}$$

We then use $\{w(s) : s \in G\}$ to perform weighted sampling, drawing for each epoch n sentence pairs without replacement. While all selection weights are very close to

[2]Scripts to generate bitext rankings and to apply both of the dynamic data selection techniques are available on `github.com/marliesvanderwees/dds-nmt`.

zero, higher ranked sentences have a noticeably higher probability of being selected than lower-ranked sentences. In practice we find that top-ranked sentences get selected in nearly each epoch, while bottom-ranked sentence pairs get selected at most once. Note that the sampled selection for any epoch is independent of selections for all other epochs.

8.4.2 Gradual fine-tuning

The second dynamic data selection technique, see Figure 8.1b, is inspired by the success of domain-specific fine-tuning (Luong and Manning, 2015; Zoph et al., 2016; Sennrich et al., 2016b; Freitag and Al-Onaizan, 2016), in which a model trained on a large general-domain bitext is trained for a few additional epochs only on small in-domain data. However, rather than training a full model on the complete bitext G, we gradually decrease the training data size, starting from G and keeping only the top n sentence pairs for the duration of η epochs, where the top n pairs are defined by their CED_s scores. Given its resemblance to fine-tuning, we refer to this variant as *gradual fine-tuning*.

During gradual fine-tuning, the selection size n is a function of epoch i:

$$n(i) = \alpha \cdot |G| \cdot \beta^{\lfloor (i-1)/\eta \rfloor}. \tag{8.5}$$

Here $0 \leq \alpha \leq 1$ is the *relative start size*, i.e., the fraction of general bitext G used for the first selection, $0 \leq \beta \leq 1$ is the *retention rate*, i.e., the fraction of data to be kept in each new selection, and $\eta \geq 1$ is the number of consecutive epochs each selected subset is used for. Note that $\lfloor i/\eta + 1 \rfloor$ indicates rounding down $i/\eta + 1$ to the nearest integer. For example, if we start with the complete bitext ($\alpha = 1$), select the top 60% ($\beta = 0.6$) every second epoch ($\eta = 2$), then we run epochs 1 and 2 with a subset of size $|G|$, epochs 3 and 4 with a subset of size $0.6 \cdot |G|$, epochs 5 and 6 with a subset of size $0.36 \cdot |G|$, and so on. For every size n, the actual selection contains the top n sentences pairs of G.

8.5 Experimental settings

We evaluate static and dynamic data selection on a German-English translation task comprising four test sets. Below we describe the MT systems and data specifications.

8.5.1 Machine translation systems

While the main aim of this chapter is to improve data selection for NMT, we also perform comparative experiments using PBMT. Our PBMT system is trained using Oister (Section 2.4.1). To create optimal PBMT systems given the available resources, we apply test-set-specific parameter tuning and LM interpolation weights optimization. Consistent with Axelrod et al., we do not vary the target-side LM between different experiments on the same test set. All n-gram models in our work are 5-gram.

For our NMT experiments we use in-house system Tardis (Section 2.4.2). We train for 16 epochs and test on the model from the last epoch. All NMT experiments are run on a single NVIDIA Titan X GPU.

8.5.2 Training and evaluation data

We evaluate all experiments on four domains: (i) EMEA medical guidelines (Tiedemann, 2009a), (ii) movie dialogues (Chapter 7) constructed from OpenSubtitles (Lison and Tiedemann, 2016), (iii) TED talks (Cettolo et al., 2012), and (iv) WMT news. For TED, we use IWSLT2010 as development set and IWSLT2011-2014 as test set, and for WMT we use newstest2013 as development set and newstest2016 as test set. We train our systems on a mixture of domains, comprising Commoncrawl, Europarl, News Commentary, EMEA, Movies, and TED. Corpus specifications are listed in Table 8.1.

Table 8.1: Data specifications with tokens counted on the German side. The WMT training corpus contains Commoncrawl, Europarl, and News Commentary but no in-domain news data.

	Train		Dev/valid		Test	
Corpus	Lines	Tokens	Lines	Tokens	Lines	Tokens
EMEA	206K	3.3M	3.9K	59K	5.8K	93K
Movies	101K	1.2M	4.5K	54K	7.1K	87K
TED	189K	3.3M	2.5K	50K	5.4K	99K
WMT	3.8M	84M	3.0K	64K	3.0K	65K
Mix	4.3M	92M	3.5K	61K	–	–

The in-domain LMs used to rank training sentences for data selection are trained on small portions of in-domain parallel data whenever available (3.3M, 1.2M and 3.3M German tokens for EMEA, Movies and TED, respectively). Since no sizable in-domain parallel text is available for WMT, we independently sample 200K sentences (3.3M German tokens or 3.5M English tokens) from the WMT monolingual News Crawl corpora. This demonstrates the applicability of data selection techniques even in cases where one lacks parallel in-domain data.

Before running data selection, we preprocess our data by tokenizing, lowercasing and removing sentences that are longer than 50 tokens or that are identified as a different language. After selection, we apply Byte-pair encoding (BPE, Sennrich et al. (2016c)) with 40K merge operations on either side of the complete mix-of-domains training bitext. For our NMT experiments we use BPE-processed corpora on both bitext sides, while for PBMT we only apply BPE to the German side. Our NMT systems use a vocabulary size of 40K on both the source and target side.

8.6 Results

Below we discuss the results of our translation experiments using static and dynamic data selection, measuring translation quality with case-insensitive untokenized BLEU.

8.6.1 Static data selection for PBMT and NMT

We first compare the effects of static data selection with n-gram LMs for both NMT and PBMT using various selection sizes. Concretely, we select the top n sentence pairs

such that the number of selected tokens $t \in \{5\%, 10\%, 20\%, 50\%\}$ of G, or $t = |I|$, the in-domain corpus size. Figure 8.2 shows translation performance measured in BLEU for our four test sets. The benefits of n-gram-based data selection for PBMT (purple circles) are confirmed: In all test sets, the selection of size $|I|$ (dotted vertical line) yields better performance than using only the in-domain data of the exact same size (purple star), and at least one of the selected subsets—often using only 5% of the complete bitext—outperforms using the complete bitext (light purple line). We also show that the informed selections are superior to random selections of the same size (purple diamonds).

In NMT, results of n-gram-based data selection (green triangles) vary: while for Movies a selection of only 10% outperforms the complete bitext (light green line), none of the selected subsets for other test sets is noticeably better than the full bitext.[3] Interestingly, the same selections of size $|I|$ that proved useful in PBMT, never beat the system that uses exactly the available in-domain data (green star), indicating that the current selections can be further improved for NMT. In all scenarios we see that NMT suffers much more from small-data settings than PBMT. Finally, the random selections (green squares) show that NMT not only needs large quantities of data, but it is also affected when the selected data is not relevant for the test data. In PBMT, both low-quantity and low-quality scenarios appear to be compensated for by the large monolingual LM on the target side.

When comparing the test sets, we observe that the impact of domain mismatch in NMT with respect to PBMT is largest for the two domains that are most distinct from the general bitext, EMEA and Movies. For WMT, both MT systems achieve very similar baseline results, but translation quality deteriorates considerably in data selection experiments, which is likely caused by the lack of in-domain data in the general bitext.

Table 8.2: NMT BLEU scores when using n-gram LMs and LSTMs for bitext ranking. Selection sizes concern the selected bitext subsets: LMs are always created from the same in-domain data.

Selection	LM type	EMEA	Movies	TED	WMT
5%	n-gram	29.8	17.4	22.6	8.1
	LSTM	**30.0**	**17.8**	22.6	**9.6**
10%	n-gram	33.0	19.6	24.5	16.6
	LSTM	33.0	**19.7**	**24.7**	**17.4**
20%	n-gram	**34.8**	19.0	25.6	21.9
	LSTM	34.5	**19.6**	**26.6**	21.9

LSTM versus n-gram Before proceeding with dynamic data selection for NMT, we test whether bitext ranking for NMT can be improved using LSTMs rather than conventional n-gram LMs. Table 8.2 shows NMT BLEU scores of a few different sizes of selected subsets created using n-gram LMs or LSTMs. While results vary among test sets and selection sizes, we observe an average improvement of 0.4 BLEU when using

[3] Validation cross-entropy converges after 10–12 epochs, never reaching the scores of the complete bitext.

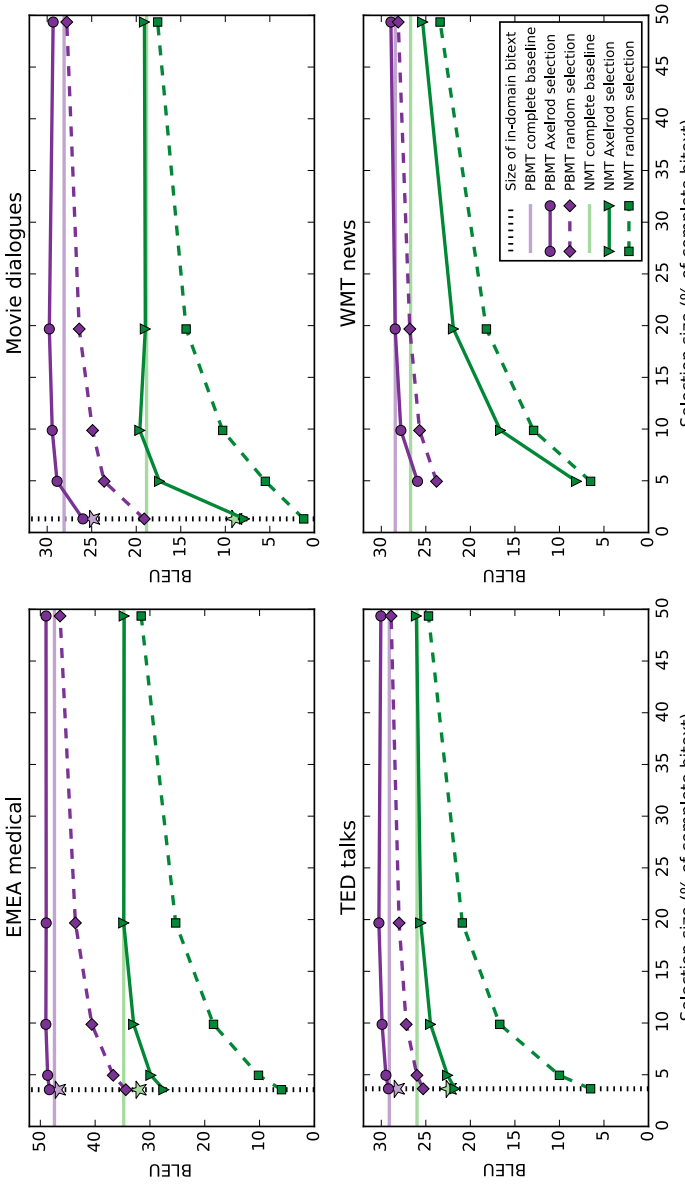


Figure 8.2: PBMT (purple) and NMT (green) German-English results of Axelrod data selection and random data selection (average of three runs) for four domains. Purple and green stars indicate BLEU scores when only the available in-domain data is used. We use selections of the in-domain size $|I|$, and 5%, 10%, 20%, and 50% of the complete bitext, which are exactly the same for PBMT and NMT.

LSTMs instead of n-gram LMs. For PBMT, similar results have been reported when replacing n-gram LMs with recurrent neural LMs (Duh et al., 2013). In all subsequent experiments we use relevance rankings computed with LSTMs instead of n-gram LMs.

8.6.2 Dynamic data selection for NMT

Equipped with a relevance ranking of sentence pairs in bitext G, we now examine two variants of dynamic data selection as described in Section 8.4.

We are interested in reducing training time while limiting the negative effect on BLEU for various domains. Therefore we report BLEU as well as the *relative training time* of each experiment. Since wall-clock times depend on other factors such as the NMT architecture and memory speed, we define training time as the total number of tokens observed while training the NMT system, i.e., the sum of tokens in the selected subsets of all epochs. We report all training times relative to the training time of our complete-bitext baseline, i.e., 4.3M tokens $\times$ 16 epochs. Note that this measure of training time corresponds closely but not exactly to the number of model updates, as the latter relies on the number of sentences, which vary in length, rather than the number of tokens in the training data. For completeness: training the 100% baseline takes 106 hours, while our fastest dynamic selection variant takes 19–21 hours. Computing CED scores takes $\sim$15 minutes when using n-gram LMs and 5–6 hours when using LSTMs.

Figure 8.3 shows BLEU scores of some selected experiments as a function of relative training time. Compared to static data selection (blue lines), our weighted sampling technique (orange triangles) yields variable results. When sampling a subset of 20% of $|G|$ from the top 50% of the ranked bitext, we obtain small improvements for TED and WMT, but small drops for EMEA and Movies. Other selection sizes (30% and 40%, not shown) give similar results lacking a consistent pattern.

By contrast, our gradual fine-tuning method performs consistently better than static selection, and even beats the general baseline in three out of four test sets. The displayed version uses settings $(\alpha = 0.5, \beta = 0.7, \eta = 2)$ and is at least as fast as static selection using 20% of the bitext, yielding up to +2.6 BLEU improvement (for WMT news) over this static version. Compared to the complete baseline, this gradual fine-tuning method improves up to +3.1 BLEU (for TED talks).

Table 8.3 provides detailed information on additional experiments using other settings. For all three test domains that are covered in the parallel data—EMEA, Movies and TED—improvements are highest when starting gradual fine-tuning with only the top 50% of the ranked bitext, which are also the fastest approaches. For WMT, which is not covered in the general bitext, adding more data clearly benefits translation quality. These findings are consistent with the static data selection patterns: using low-ranked sentences on top of the most relevant selection does not improve translation performance for any domain except WMT news.

Finally, we compare our data selection experiments to domain-specific fine-tuning (light blue stars in Figure 8.3), which is the current state-of-the-art for domain adaptation in NMT. To this end, we first train a model on the complete bitext, and then train for twelve additional epochs on available in-domain data, using an initial learning rate of 1 which is halved every epoch. Depending on the test set, this approach yields +2.5–4.4 BLEU improvements over our baselines, however it does not speed up training

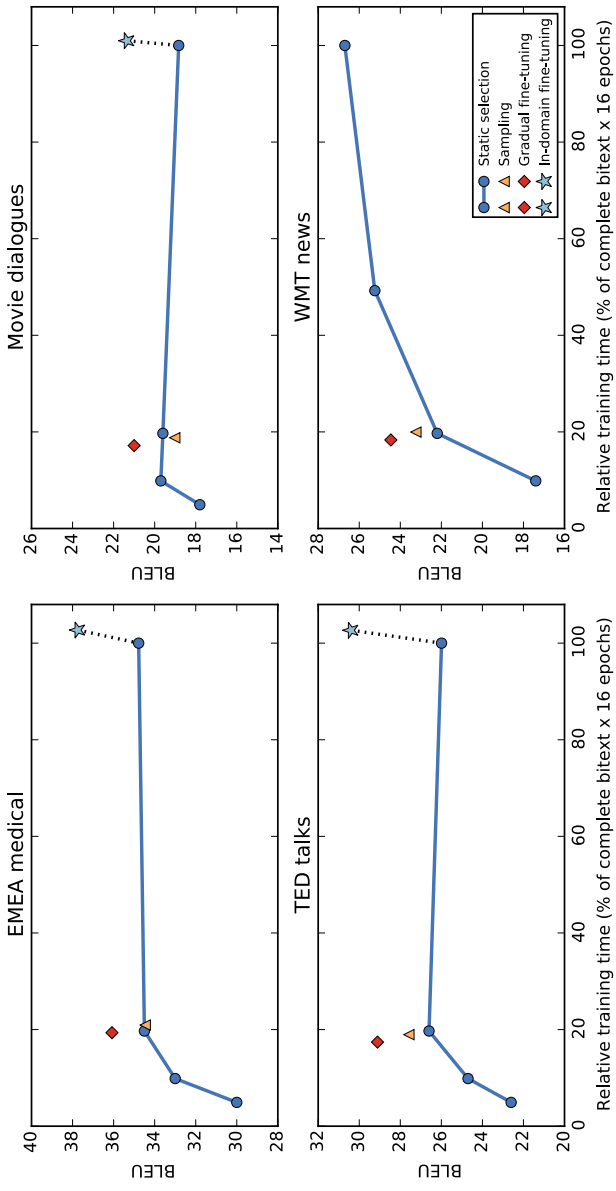


Figure 8.3: Selected German-English translation results of dynamic data selection methods (orange and red markers) compared to conventional static data selection (blue circles). *Relative training time* equals the total number of training tokens relative to the complete baseline, which takes 106 hours to train and is represented by the rightmost blue circle. Note that no parallel in-domain data is available for WMT news. All y-axes are scaled equally for easy comparison of BLEU differences across domains.

Table 8.3: German-English BLEU results of various gradual fine-tuning experiments sorted by relative training time. Indicated improvements are with respect to static selection using 20% of the bitext, and highest scores per test set are bold-faced. Results from the first experiment are also shown in Figure 8.3.

Experiment			Relative training time	BLEU			
Start size	Retention rate β	Decrease every		EMEA	Movies	TED	WMT
Static selection top 20%			20%	34.5	19.6	26.6	21.9
50% ($\alpha = 0.5$)	0.7	$\eta = 2$ epochs	18–20%	**36.1 (+1.6)**	21.0 (+1.4)	**29.1 (+2.5)**	24.5 (+2.6)
50% ($\alpha = 0.5$)	0.5	$\eta = 4$ epochs	21–23%	36.0 (+1.5)	**21.2 (+1.6)**	29.0 (+2.4)	25.0 (+3.1)
50% ($\alpha = 0.5$)	0.6	$\eta = 4$ epochs	25–27%	35.6 (+1.1)	21.0 (+1.4)	28.5 (+1.9)	25.1 (+3.2)
100% ($\alpha = 1$)	0.6	$\eta = 2$ epochs	29–31%	35.5 (+1.0)	21.1 (+1.5)	29.0 (+2.4)	25.6 (+3.7)
100% ($\alpha = 1$)	0.7	$\eta = 2$ epochs	37–39%	35.9 (+1.4)	20.4 (+0.8)	28.2 (+1.6)	25.8 (+3.9)
100% ($\alpha = 1$)	0.9	$\eta = 1$ epoch	50–52%	35.4 (+0.9)	19.6 ($\pm$0.0)	27.4 (+0.8)	**26.1 (+4.2)**
Complete bitext baseline			100%	34.8	18.8	26.0	26.7
Gold: fine-tuning on in-domain data			101–103%	37.7	21.3	30.4	–

and requires a parallel in-domain text which may not be available, e.g., for WMT. While none of our data selection experiments outperforms domain-specific fine-tuning, we obtain competitive translation quality requiring only 20% of the training time. In additional experiments we found that in-domain fine-tuning on top of our selection approaches does not yield further improvements.

8.7 Further analysis

In this section we conduct a few additional experiments and analyses. We restrict ourselves to one parameter setting per selection approach: static selection and sampling with 20% of the data, and gradual fine-tuning using $(\alpha = 0.5, \beta = 0.7, \eta = 2)$. All have very similar training times.

First, we hypothesize that dynamic data selection works well because more unique sentence pairs are observed during training, and it therefore increases coverage with respect to static data selection. To verify this, we measure for each test set the number of unseen source word types in the training data for different selection methods. Figure 8.4 shows that the average number of unseen word types is indeed reduced noticeably in both of the dynamic selection techniques, being much closer to the complete bitext baseline than to static selection. Note that all methods use the same vocabulary during training.

Next, following the static data selection experiments in Section 8.6.1, we examine how well dynamic data selection performs using random selections. To this end, we repeat all techniques using a bitext which is ranked randomly rather than by its relevance to the test sets. The results in Table 8.4 show that the bitext ranking plays a crucial role in the success of data selection. However, the results also show that *even* in the absence of an appropriate bitext ranking, dynamic data selection—and in particular gradual fine-tuning—is still superior to static data selection. We explain this result as follows: Compared to static selection, both sampling and gradual fine-tuning have better coverage due to their improved exploration of the data. However, sampling also suffers from a surprise effect of observing new data in every epoch. Gradual fine-tuning on the other hand gradually improves learning on a subset of the selected data, suggesting that repetition across epochs has a positive effect on translation quality.

Table 8.4: BLEU scores of data selection using relevance versus random ranking of the bitext. Gradual fine-tuning uses $(\alpha = 0.5, \beta = 0.7, \eta = 2)$, with relative training times of 18–20%.

Ranking	Method	EMEA	Movies	TED	WMT
Relevance	Gradual FT	**36.1**	**21.0**	**29.1**	**24.5**
	Sampling 20%	34.5	19.0	27.6	23.2
	Static 20%	34.5	19.6	26.6	21.9
Random	Gradual FT	**29.2**	**16.1**	**23.2**	**21.3**
	Sampling 20%	26.7	14.4	22.0	19.8
	Static 20%	25.3	14.4	20.9	18.2

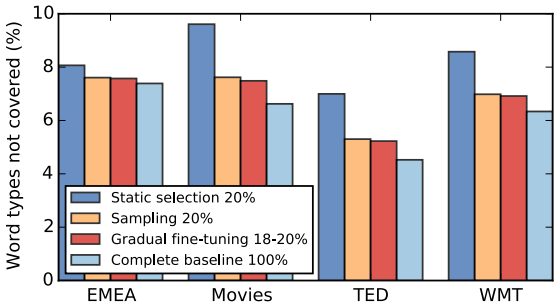


Figure 8.4: Test set source words not covered in the training data of different data selection methods.

One could expect that changing the data during training results in volatile training behavior. To test this, we inspect the cross-entropy of our development sets after every training epoch. Figure 8.5 shows these results for TED. Clearly, static data selection converges most steadily. However, both dynamic selection techniques eventually converge to a lower cross-entropy value which is reflected by higher translation quality of the test set. We observe very similar behavior for the other test sets.

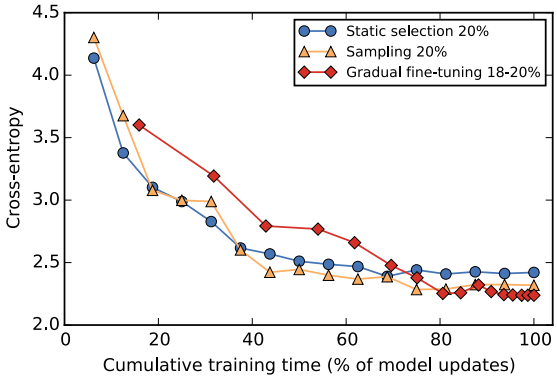


Figure 8.5: German-English cross-entropy of the TED dev set as a function of training time. Each data point represents a completed training epoch.

By its nature, our gradual fine-tuning technique uses training epochs of different sizes, and therefore also implicitly differs from other methods in its parameter optimization behavior. Since we decrease both the training data size and the SGD learning rate after finishing complete training epochs, we automatically decay the learning rate at decreasing time intervals. We therefore study how this approach is affected when we (i) decay the learning rate after a fixed number of updates (i.e., the same as in static data selection) rather than per epoch, or (ii) keep the learning rate fixed. In the first scenario, we observe that translation performance drops with –1.1–2.0 BLEU. When keeping a fixed learning rate, BLEU scores hardly change or even improve, indicating that the implicit change in search behavior may contribute to the success of gradual fine-tuning.

8.8 Conclusions

With the recent increase in popularity of NMT, we explored in this chapter *to what extent* and *how* NMT can benefit from data selection. The potential benefits of applying data selection are twofold; data selection can improve translation quality by discarding data that is not relevant for the domain of interest, and it can substantially speed up training times, which are known to be long for NMT training. When used for domain adaptation, automatic data selection has the advantage that no prior knowledge is needed about which sentence pairs or subcorpora are most relevant for the task at hand. Data selection is therefore a method that meets one of the main goals of this thesis: developing adaptation approaches that do not rely on the availability of domain or subcorpus information. Motivated by good results of data selection in PBMT, we asked:

RQ6 *To what extent and how can we successfully apply data selection for domain adaptation in NMT?*

To answer this question, we first investigated the effects of an existing data selection method on both PBMT and NMT. On a German-English translation task covering four domains, we observed that the data selection method performs consistently well for PBMT but yields unreliable results for NMT.

Next, in order to make data selection a more effective strategy for NMT, we introduced *dynamic data selection*, which entails varying the selected subset of training data between different training epochs. We explored two techniques of dynamic data selection, both of which employ a domain-relevance ranked bitext: *sampling*, in which we sample sentence pairs from a weighted distribution based on each sentence's relevance to the domain of interest, and *gradual fine-tuning*, in which we gradually reduce the training data size, keeping the decreasing top n most relevant sentence pairs for training.

While results using our sampling approach vary, we found that our gradual fine-tuning technique improves consistently over conventional static data selection (up to +2.6 BLEU) and over a high-resource general baseline (up to +3.1 BLEU). Moreover, gradual fine-tuning approximates in-domain fine-tuning using only $\sim$20% of the training time, even when no parallel in-domain data is available. In our additional analysis we found that potential explanations for the success of gradual fine-tuning are (i) the increased vocabulary coverage, and (ii) the implicit change in search behavior resulting from the gradually decreasing size of the training data.

9 Conclusions

In this thesis we have shed light on domain adaptation for machine translation (MT), and in particular on the effects of various fine-grained aspects that together make up a domain. This research is motivated by the fact that despite domain adaptation being a well-studied research topic, previous research ignores three major issues: first, the notion of a *domain* is not unambiguously defined, often referring to provenance while disregarding that different domains may vary at many other levels. Second, research in MT is mostly driven by relatively formal, editorially controlled, translation tasks, while *informal* genres are much harder to translate. Third, most existing adaptation approaches depend on the availability of provenance labels in both the training and the evaluation data.

This thesis covers three research themes. In the first theme we have clarified the concept of a domain and disentangled its aspects *genre* and *topic*. We found that genre differences pose a larger challenge to PBMT than topic differences and analyzed why PBMT struggles with genre differences. Next, we introduced two genre adaptation approaches that do not, or at most partially, depend on the availability of domain or provenance information of the training or the test data.

In the second research theme we analyzed PBMT for *informal* genres. While most research is driven by translation tasks with standardized language use, informal data is known to suffer from poor translation quality. In this thesis, we considered in total six types of informal data (SMS messages, chat messages, telephone conversations, weblogs, user comments, and fictional dialogues), and we analyzed which challenges they pose to PBMT and what are promising strategies to improve translation quality.

In the third research theme we addressed domain adaptation for the new paradigm of NMT. Concretely, we investigated whether an existing adaptation approach based on data selection can be applied to NMT. After observing that this method performs consistently well for PBMT but yields unreliable results for NMT, we introduced a novel data selection approach which achieves solid BLEU improvements and makes data selection beneficial for domain adaptation in NMT.

9.1 Main findings

In this section, we revisite and answer our research questions and we summarize the main findings within each of our three research themes.

Part I. The role of genres in phrase-based machine translation

Within the first research theme we have answered three main research questions. First, we addressed the fact that within the active research field of domain adaptation for PBMT, there is little consensus on what exactly constitutes a domain. Typically, different domains correspond to different subcorpora or data sets. Since most previous work on domain adaptation in PBMT uses in-domain and out-of-domain data that differ on both the *topic* and the *genre* level, it is unclear whether the proposed solutions help deal with topic or genre differences. We have disentangled these two concepts and studied their respective impact on PBMT by asking:

RQ1 *What impact do genre and topic differences have on PBMT quality?*

To answer this question, we first provided clear definitions to the concepts of topic and genre: topic refers to the subject of a text and is mostly characterized by specific vocabulary or word senses, while genre concerns functional and stylistic properties and is characterized by aspects such as writing style, stopwords, and register. Using these definitions, we introduced a new annotated Arabic-English benchmark set with controlled topic and genre distributions covering newswire (NW) and user-generated (UG) documents in five topics.

When translating this data using a genre-balanced PBMT system, we found that translation performance in BLEU fluctuates much more between the two genres than across topics and that UG is translated with shorter phrases than NW. To explain these observations, we introduced new quantitative metrics measuring *source phrase recall* and *source-target phrase pair recall*, representing the source and source-target coverage of our models. We found that poor model coverage—both mono- and bilingual—is an important reason for the low PBMT quality on UG data. Our manual qualitative analysis at the word level also suggests that source coverage could benefit from text normalization, paraphrasing, or efforts to increase the amount of relevant training data.

Next, we further explored the impact of different genres on PBMT quality. To this end, we extended our research to four genres (colloquial, editorial, news, and speech) for four language pairs (Arabic-English, Bulgarian-English, Chinese-English, and Persian-English), for which we harvested parallel training, development and testing corpora from the web, and then asked:

RQ2 *Is the observed impact of genre differences on PBMT consistent among various language pairs and data settings?*

We found that variations in BLEU are to a large extent language-specific, we also observed moderate to strong positive correlations between genre-specific proportions of the training data and BLEU scores. This indicates that differences in PBMT performance can to a certain degree be explained by the amount of available training data. Moreover, a few genre-specific patterns generalize across language pairs. For example, news is always relatively easy to translate, whereas colloquial and speech documents achieve relatively low BLEU scores for all language pairs.

After studying the impact of genre differences on PBMT, we moved to genre adaptation and asked:

RQ3 *How can we adapt PBMT systems to different genres without relying on explicit corpus labels?*

We explored two approaches to improve genre adaptation for PBMT. First, we exploited our new corpora to train genre-specific rather than genre-agnostic PBMT systems and obtained significant BLEU improvements when manually routing each test document to the most appropriate genre-specific system. However, since we do not want to rely on manual labels, we incorporated document-level genre classifiers into an end-to-end PBMT pipeline and showed that we can improve translation quality over the genre-agnostic baseline system without having to know a test document's genre in advance. We also found that misclassification does not always lead to deterioration of translation quality, but rather benefits some documents.

For our second approach to genre adaptation we improved upon an existing adaptation framework by replacing its dependency on subcorpus information with automatic indicators of genre. We explored two types of features that can be automatically extracted from the parallel training and evaluation corpora: genre-revealing features inspired by previous findings in the text classification literature, and latent Dirichlet allocation (LDA) features. While both feature sets improved PBMT for most of our test genres, we found that combining the feature sets yields the largest BLEU improvements, indicating that the proposed genre and LDA features are to some extent complementary.

Part II. Translating and analyzing informal language

In our second research theme we addressed two main research questions, both dealing with translating informal genres. First, we examined what exactly makes translating user-generated (UG) text such as found on social media and web forums a difficult task for PBMT. We collected and analyzed five different types of UG data: SMS messages, chat messages, manual transcripts of phone conversations, weblogs, and readers' comments to news articles, and we asked:

RQ4 *How is translation quality of PBMT influenced by different types of user-generated text?*

To answer this question, we compared our five UG benchmarks both quantitatively and qualitatively. Our quantitive analysis included measuring translation quality, average translation phrase length, and model coverage of source phrases, target phrases and phrase pairs. We also compared for each UG data set the distribution of errors caused by unseen source words, unseen target words, or suboptimal scoring of possible translation candidates. Our results show among others that (i) UG data is translated with shorter source phrases than news, (ii) UG translation model coverage deteriorates substantially for longer phrases, and (iii) phrase-pair OOVs pose a bigger challenge to UG translation tasks than source OOVs.

We also qualitatively analyzed PBMT errors for UG using sentence-level error annotations. In this analysis we found that common issues in UG data include (i) OOVs due to misspellings or Arabic dialectal forms, (ii) lexical choices that do not reflect colloquial formulations, (iii) phrasal idioms being translated word by word, and (iv) omitted first person pronouns in Chinese SMS and chat.

In addition, we performed all of our analyses for two news benchmarks, allowing us to compare PBMT performance and errors for each of the UG data sets to those observed for a more formal genre. We found that the SMS and chat benchmarks are the most dissimilar to news at all the analyzed levels. Errors in other types of UG are often more similar to news errors than to those in SMS and chat messages.

Our second research question in this theme deals with the informal genre of dialogues. Dialogues involve, by definition, multiple *speakers*, often with different *genders*, who likely have varying intentions and language use. Dialogues are therefore characterized by functional actions known as *dialogue acts* and utterances with varying *register* levels. Since the impact of these dialogue-specific aspects on PBMT has hardly been studied, we investigated the effect of four dialogue-specific aspects and asked:

RQ5 *What impact do dialogue-specific aspects have on translation quality?*

First, we created movie-dialogue benchmarks for five language pairs in which utterances are annotated with dialogue acts, speakers, speaker gender, and register. Using this new resource we found that BLEU fluctuations for speaker gender, dialogue acts, and register levels are always significantly larger than fluctuations of randomly shuffled data subsets. As for speaker variations, we found that on average 18.6% of the movies show larger fluctuations than expected at random, suggesting that for this subset of movies adaptation towards speakers may benefit PBMT quality.

We found that the register level has a significant impact on translation quality, with translations of vulgar utterances being of substantially lower quality than neutral or colloquial utterances. Similarly we found large variations in translation quality between different dialogue acts, although we did not detect a consistent pattern between different languages, e.g., questions can be more difficult to translate than exclamations for one language pair, while the reverse is true for another language pair.

These observations suggest that conversational PBMT may benefit from adaptation at fine-grained levels, which we tested in a series of adaptation experiments towards dialogue acts and register levels. We confirmed that apart from domain adaptation, adaptation to other variables should be considered to improve PBMT quality for dialogues. While our analyses are carried out on fictional dialogues, we believe that our findings generalize at least partially to real dialogues, and are thus valuable for advancing conversational MT.

Part III. Domain adaptation for neural machine translation

In the third and final research theme we addressed domain adaptation for the new MT paradigm of NMT. Concretely, we explored *to what extent* and *how* NMT can benefit from data selection, asking:

RQ6 *To what extent and how can we successfully apply data selection for domain adaptation in NMT?*

To answer this question, we first investigated the effects of an existing, widely used data selection method (Axelrod et al., 2011) on both PBMT and NMT. On a German-English translation task covering four test domains, we observed that, while consistently

performing well for PBMT, this method yields unreliable results for NMT, resulting in substantial drops in BLEU for small selection sizes.

Next, in order to make data selection a more effective strategy for domain adaptation in NMT, we introduced *dynamic data selection* for NMT, which entails varying the selected subset of training data between different training epochs. We explored two techniques of dynamic data selection, both of which employ a domain-relevance ranked bitext: (i) *sampling*, in which we sample sentence pairs from a weighted distribution based on each sentence's relevance to the domain of interest, and (ii) *gradual fine-tuning*, in which we gradually reduce the training data size, keeping the decreasing top n most relevant sentence pairs for training.

While results using our sampling approach vary, we found that our gradual fine-tuning technique improves consistently over conventional static data selection (up to +2.6 BLEU) and over a high-resource general baseline (up to +3.1 BLEU). Moreover, gradual fine-tuning approximates in-domain fine-tuning using only $\sim$20% of the training time, even when no parallel in-domain data is available. In our additional analysis we found that potential explanations for the success of gradual fine-tuning are (i) the increased vocabulary coverage, and (ii) the implicit change in search behavior resulting from the gradually decreasing size of the training data.

9.2 Future work

In this thesis we have shed light on domain adaptation for machine translation, and in particular on the effects of various fine-grained aspects that together make up a domain: genres, topics, speakers, gender, dialogue acts, and register levels. In addition, we have presented a number of adaptation approaches that do not, or at most partially, depend on the availability of domain or provenance information of the training or the test data. Both of these research directions serve the ultimate goal of *dynamic adaptation* for MT, a scenario we discuss below in Section 9.2.1.

While most of the work in this thesis concerns statistical PBMT, research is currently mostly driven by the newly developed paradigm of neural MT (NMT). We therefore dedicate Section 9.2.2 to what NMT can learn from findings in PBMT when dealing with domain adaptation scenarios.

9.2.1 Dynamic and fine-grained adaptation for MT

In an ideal scenario, the entire adaptation process would take place at translation time and the MT system is adjusted for each incoming translation task. This type of dynamic adaptation can be realized at many different levels of granularity. Besides adapting an MT system to a complete test set (see Chapters 5 and 8), to a document (see Chapters 3 and 4), or to different sentences (see Chapter 7), one could for example consider adaptation at the sub-document level. For example, previous work has demonstrated that changes in a document's discourse structure can be automatically detected using features observable in raw text (Friedrich and Pinkal, 2015a,b; Friedrich et al., 2016), an interesting potential direction for MT adaptation. Alternatively, it might be beneficial to look beyond individual sentences and take into account each sentence's direct context,

e.g., previous and following sentences. This is particularly promising for conversational genres, where subsequent utterances can vary at many different levels but may capture information about other utterances.

A related interesting direction for future research is investigating how to simultaneaously adapt to different aspects at once, for example to the genre and topic of a document and also to the register level of individual sentences. Combining topic and provenance adaptation has been done by Hasler et al. (2014b), but combining other aspects has not been described before in the research literature. For PBMT, combined adaptation can be achieved for instance by incorporating relevance weights (Chapter 5) representing different aspects of a domain, while for NMT an encouraging strategy would be to extend our gradual fine-tuning method (Chapter 8), for example by first fine-tuning towards a document's genre and then further fine-tuning towards each sentence's register level, or by combining model ensembles.

The above mentioned methods adapt MT systems by re-estimating parameter values. However, we showed in Chapters 3 and 6 that translation errors for most informal genres are more attributable to poor model coverage than to suboptimal scoring of existing translation candidates. This can be solved by augmenting the training data with genre-specific translation examples, i.e., phrase or sentence pairs. However, data augmentation in turn increases model coverage and thus the complexity of the decoding search space, a problem which can be addressed using data selection. While data augmentation and data selection strategies seem complementary, they both aim to yield only highly relevant training data, and preferably as much of it as possible. It is therefore interesting to explore methods to simultaneously synthesize relevant data *and* discard irrelevant data. Unfortunately, adaptation at the data level typically takes place at training time, hence a system has to be fully re-trained for successful adaptation. Relevant future work therefore also includes addressing training efficiency when shifting to new domains.

Another strategy that can be applied to both PBMT and NMT is adaptation using pre-translation text classifiers. While the genre classifiers used in Chapter 4 are all crisp classifiers, i.e., they assign a single label per test document, a different strategy is to apply fuzzy classification, in which different MT models can be interpolated, e.g., linearly or log-linearly in PBMT or using ensembles in NMT, using classifier output weights. Such a setting might particularly benefit documents with low classification confidence. Additionally, it is worth investigating data clustering strategies, such that no prior information is needed about domains or genres in the training data. This strategy has the additional advantage to possibly scale to an arbitrary large number of domain aspects since no pre-defined categories are needed.

Finally, dynamic adaptation for PBMT also implies eliminating the need of a development set that is representative of the test set's genre or topic distribution.

9.2.2 What NMT adaptation can learn from PBMT adaptation

While most of the work in this thesis has been applied to PBMT, the novel promising paradigm to automatic translation is that of neural MT. Research in MT is shifting towards NMT, and NMT achieves state-of-the-art performance for an increasing number of language pairs and translation tasks (Bojar et al., 2015, 2016). It is thus likely that future work on adaptation for MT mainly involves adaptation for NMT.

Despite these promising results, NMT is still in its infancy and faces a number of serious challenges (Koehn and Knowles, 2017), one of which is its poor performance in domain-mismatch scenarios. To date, this has been addressed by domain adaptation approaches which almost all follow a similar setup called *fine-tuning*: after training a general-domain model, a few additional training iterations are completed using in-domain data. While fine-tuning has the advantage that it can be applied quickly if a general-domain system is available, it is worth designing other methods for domain adaptation in NMT.

Since domain adaptation for PBMT has been a very active field for more than a decade, there are a number of things NMT adaptation can learn from PBMT adaptation. First, an important lesson learned from this thesis is that PBMT performance can benefit from adaptation to fine-grained domain aspects rather than only to provenance-based domains, which is consistent with findings in previous work on topic adaptation. A logical next step in NMT adaptation research is to move away from provenance-based domain definitions and examine other aspects as indicators for adaptation.

Second, many of the successful adaptation methods for PBMT exploit some sort of feature augmentation, in which additional information about the domain or genre of interest is represented with one or more additional features. In NMT, such an approach has been applied occasionally, for example to control politeness of German output translations, i.e., du versus Sie (Sennrich et al., 2016a), or to capture hard domain labels during adaptation (Chu et al., 2017). This leaves room for an extensive exploration of other additional features or side constraints to use, preferably features that can be inferred from raw text without requiring domain knowledge.

Third, an implicit consequence of adapting—or one could say, overfitting—an MT system to a specific domain is that it suffers from degraded translation performance on out-of-domain data, making an adapted system unsuitable for any other domain. In PBMT this issue has been addressed by means of model interpolation or combination (Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Bisazza et al., 2011) or using pre-trained classifiers (Chapter 4, Wang et al. (2012)) such that one can achieve reasonable performance on various domains. While this issue has recently been addressed by Dakwale and Monz (2017) with a modified fine-tuning approach, it is interesting to also explore model combination or adaptation using classifiers.

Finally, many aspects of NMT are still considered as a *black box*, and future analyses have to provide insights into the successes and weaknesses of NMT. In the context of domain adaptation for PBMT, previous work has analyzed in which stage of the pipeline the available in-domain data can best be used (Duh et al., 2010), how in-domain and out-of-domain models can best be combined (Bisazza et al., 2011), or whether it is more promising to improve either phrase extraction or scoring (Haddow and Koehn, 2012). While NMT architectures do not consist of separate model components, it is still worth investigating what impact in-domain data, for example varying in amount or quality, has on translation performance, albeit only to ignite new research directions.

Bibliography

A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, 2011. (Cited on pages 1, 5, 31, 39, 94, 115, 116, 117, 118, 119, 120, 121, and 134.)

A. Axelrod, P. Resnik, X. He, and M. Ostendorf. Data selection with fewer words. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 58–65, 2015. (Cited on pages 31 and 117.)

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. (Cited on pages 18, 20, and 115.)

R. E. Banchs. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 203–207, 2012. (Cited on pages 99 and 100.)

R. E. Banchs and M. R. Costa-Jussà. A semantic feature for statistical machine translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 126–134, 2011. (Cited on pages 33 and 69.)

P. Banerjee, J. Du, B. Li, S. Kumar Naskar, A. Way, and J. van Genabith. Combining multi-domain statistical machine translation models using automatic classifiers. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, 2010. (Cited on pages 32 and 46.)

P. Banerjee, S. K. Naskar, J. Roturier, A. Way, and J. van Genabith. Domain adaptation in statistical machine translation of user-forum data using component level mixture modelling. In *Proceedings of the XIII Machine Translation Summit*, pages 285–292, 2011. (Cited on pages 1 and 85.)

P. Banerjee, S. K. Naskar, J. Roturier, A. Way, and J. van Genabith. Domain adaptation in SMT of user-generated forum content guided by OOV word reduction: Normalization and/or supplementary data. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 169–176, 2012. (Cited on page 85.)

S. Banerjee and A. Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, pages 65–72, 2005. (Cited on page 83.)

R. Bawden. Machine translation, it's a question of style, innit? The case of english tag questions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. (Cited on page 100.)

R. Bawden, G. Wisniewski, and H. Maynard. Investigating gender adaptation for speech translation. In *23ème Conférence sur le Traitement Automatique des Langues Naturelles*, 2016. (Cited on pages 99 and 108.)

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. (Cited on page 30.)

A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, A. S. Kehler, and R. L. Mercer. Language translation apparatus and method using context-based translation models, 1996. US Patent 5,510,981. (Cited on page 14.)

N. Bertoldi and M. Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, 2009. (Cited on pages 12 and 32.)

N. Bertoldi, M. Cettolo, and M. Federico. Statistical machine translation of texts with misspelled words. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 412–419, 2010. (Cited on pages 40, 85, and 96.)

A. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406, 1946. (Cited on page 71.)

A. Bisazza and M. Federico. Cutting the long tail: Hybrid language models for translation style adaptation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 439–448, 2012. (Cited on page 35.)

A. Bisazza and M. Federico. A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational linguistics*, 2016. (Cited on page 15.)

A. Bisazza, N. Ruiz, and M. Federico. Fill-up versus interpolation methods for phrase-based SMT adaptation. In *Proceedings of the 8th International Workshop on Spoken Language Translation*, pages 136–143, 2011. (Cited on pages 1, 32, 51, 85, and 137.)

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003. (Cited on pages 35 and 69.)

O. Bojar. Analyzing error types in english-czech machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:63–76, 2011. (Cited on page 85.)

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, 2013. (Cited on pages 11 and 12.)

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, et al. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, 2014. (Cited on pages 11 and 12.)

O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, et al. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, 2015. (Cited on pages 11 and 136.)

O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, et al. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation*, pages 131–198, 2016. (Cited on pages 11, 115, and 136.)

S. Brennan and C. B. Lockridge. Computer-mediated communication: A cognitive science approach. In *Encyclopedia of language and linguistics,*, pages 775–780. Elsevier Ltd., Oxford, UK, 2006. (Cited on page 98.)

P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993. (Cited on page 13.)

H. Bunt. Conversational principles in question-answer dialogues. In *Zur Theorie der Frage*, pages 119–141. Narr Verlag, 1979. (Cited on page 97.)

C. Callison-Burch, P. Koehn, and M. Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, 2006. (Cited on page 40.)

M. Carpuat and M. Simard. The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, 2012. (Cited on page 78.)

M. Carpuat, C. Goutte, and G. Foster. Linear mixture models for robust machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 499–509, 2014. (Cited on page 32.)

J. Carrera, O. Beregovaya, and A. Yanishevsky. Machine translation for cross-language social media, 2009. (Cited on page 85.)

S. Carter. *Exploration and exploitation of multilingual data for statistical machine translation*. PhD thesis, University of Amsterdam, 2012. (Cited on page 45.)

M. Cettolo, C. Girardi, and M. Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268, 2012. (Cited on page 122.)

B. Chen and F. Huang. Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 314–323, 2016. (Cited on pages 31 and 117.)

B. Chen, R. Kuhn, and G. Foster. Vector space model for adaptation in statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1285–1293, 2013. (Cited on pages 1, 33, 39, 68, 70, 71, 72, 75, 77, and 94.)

B. Chen, R. Kuhn, and G. Foster. A comparison of mixture and vector space techniques for translation model adaptation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 124–138, 2014. (Cited on pages 33 and 69.)

S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13:359–394, 1999. (Cited on pages 15 and 24.)

D. Chiang. Hierarchical phrase-based translation. *computational linguistics*, 33(2):201–228, 2007. (Cited on page 11.)

K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of the Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014a. (Cited on page 23.)

K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014b. (Cited on pages 18 and 115.)

C. Chu, R. Dabre, and S. Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 385–391, 2017. (Cited on pages 117 and 137.)

J. Civera and A. Juan. Domain adaptation in statistical machine translation with mixture modelling. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180, 2007. (Cited on

page 33.)
J. H. Clark, A. Lavie, and C. Dyer. One system, many domains: Open-domain statistical machine translation via feature augmentation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, 2012. (Cited on page 33.)
R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011. (Cited on page 18.)
Â. Costa, W. Ling, T. Luís, R. Correia, and L. Coheur. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161, 2015. (Cited on page 85.)
M. R. Costa-jussà and R. E. Banchs. A vector-space dynamic feature for phrase-based statistical machine translation. *J Intell Inf Syst*, 37(2):139–154, 2011. (Cited on pages 1, 33, and 69.)
J. Crego, J. Kim, G. Klein, A. Rebollo, K. Yang, J. Senellart, E. Akhanov, P. Brunelle, A. Coquard, Y. Deng, et al. SYSTRAN's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*, 2016. (Cited on page 118.)
H. Cuong and K. Sima'an. Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939, 2014a. (Cited on page 31.)
H. Cuong and K. Sima'an. Latent domain phrase-based models for adaptation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 566–576, 2014b. (Cited on page 33.)
H. Cuong and K. Sima'an. Latent domain word alignment for heterogeneous corpora. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 398–408, 2015. (Cited on page 33.)
H. Cuong, K. Sima'an, and I. Titov. Adapting to all domains at once: Rewarding domain invariance in SMT. *Transactions of the Association for Computational Linguistics*, 4:99–112, 2016. (Cited on page 33.)
P. Dakwale and C. Monz. Fine-tuning for neural machine translation with limited degradation across in- and out-of-domain data. In *Proceedings of the XVI Machine Translation Summit*, 2017. (Cited on pages 117 and 137.)
C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, 2011. (Cited on pages 98, 99, 100, and 108.)
H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 256–263, 2007. (Cited on page 33.)
H. Daumé III and J. Jagarlamudi. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, 2011. (Cited on page 32.)
S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990. (Cited on page 35.)
A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. (Cited on page 13.)
N. Dewdney, C. VanEss-Dykema, and R. MacMillan. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, 2001. (Cited on pages 45 and 57.)
S. Dose. Flipping the script: A corpus of american television series (CATS) for corpus-based language learning and teaching. *Corpus Linguistics and Variation in English: Focus on Non-native English*, 13, 2013. (Cited on page 98.)
K. Duh, K. Sudoh, and H. Tsukada. Analysis of translation model adaptation in statistical machine translation. In *Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 243–250, 2010. (Cited on pages 33, 85, and 137.)
K. Duh, G. Neubig, K. Sudoh, and H. Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 678–683, 2013. (Cited on pages 31, 117, and 125.)
M. Eck, S. Vogel, and A. Waibel. Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *Proceedings of the 2005 International Workshop on Spoken Language Translation*, pages 61–67, 2005. (Cited on page 117.)
E. S. Edgington. Approximate randomization tests. *The Journal of Psychology*, 72(2):143–149, 1969. (Cited on page 105.)

S. Eetemadi, W. Lewis, K. Toutanova, and H. Radha. Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4):189–223, 2015. (Cited on page 117.)

V. Eidelman, J. Boyd-Graber, and P. Resnik. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 115–119, 2012. (Cited on pages 35, 70, and 74.)

D. Elliott, A. Hartley, and E. Atwell. A fluency error categorization scheme to guide automated machine translation evaluation. In *Conference of the Association for Machine Translation in the Americas*, pages 64–73, 2004. (Cited on page 85.)

M. Fadaee, A. Bisazza, and C. Monz. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017. (Cited on pages 116 and 119.)

M. A. Farajian, M. Turchi, M. Negri, N. Bertoldi, and M. Federico. Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 280–284, 2017. (Cited on page 117.)

R. Fernández. Dialogue. In *The Oxford Handbook of Computational Linguistics (2 ed.)*. Oxford University Press, Oxford, 2014. (Cited on pages 97 and 111.)

A. Finn and N. Kushmerick. Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57:1506–1518, 2006. (Cited on pages 46 and 57.)

A. Font Llitjós, J. G. Carbonell, and A. Lavie. A framework for interactive and automatic refinement of transfer-based machine translation. In *Proceedings of the Conference of the European Association for Machine Translation*, pages 87–96, 2005. (Cited on page 85.)

P. Forchini. Spontaneity reloaded: American face-to-face and movie conversation compared. In *Corpus Linguistics*, 2009. (Cited on page 98.)

P. Forchini. *Movie language revisited. Evidence from multi-dimensional analysis and corpora*. Peter Lang, Internationaler Verlag der Wissenschaften, 2012. (Cited on page 98.)

P. Forner, A. Peñas, E. Agirre, I. Alegria, C. Forăscu, N. Moreau, P. Osenova, P. Prokopidis, P. Rocha, B. Sacaleanu, et al. Overview of the clef 2008 multilingual question answering track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 262–295, 2008. (Cited on page 100.)

G. Foster and R. Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, 2007. (Cited on pages 1, 32, and 137.)

G. Foster, C. Goutte, and R. Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, 2010. (Cited on pages 1, 33, 39, 68, and 94.)

M. Freitag and Y. Al-Onaizan. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*, 2016. (Cited on pages 117 and 121.)

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer, Berlin, 2001. (Cited on page 32.)

A. Friedrich and M. Pinkal. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, 2015a. (Cited on page 135.)

A. Friedrich and M. Pinkal. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1272–1281, 2015b. (Cited on page 135.)

A. Friedrich, A. Palmer, and M. Pinkal. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1757–1768, 2016. (Cited on page 135.)

J. Gao, J. Goodman, M. Li, and K.-F. Lee. Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing*, 1(1):3–33, 2002. (Cited on pages 31 and 117.)

G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161, 2012. (Cited on pages 115 and 117.)

F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000. (Cited on page 19.)

H. Ghader and C. Monz. What does attention in neural machine translation pay attention to? In *Proceedings of the 9th International Joint Conference on Natural Language Processing*, 2017. (Cited on page 21.)

D. Giampiccolo, P. Forner, J. Herrera, A. Peñas, C. Ayache, C. Forascu, V. Jijkoun, P. Osenova, P. Rocha, B. Sacaleanu, et al. Overview of the clef 2007 multilingual question answering track. In *Workshop of the*

Cross-Language Evaluation Forum for European Languages, pages 200–236, 2007. (Cited on page 100.)
J. Giménez and L. Màrquez. Towards heterogeneous automatic mt error analysis. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1894–1901, 2008. (Cited on page 85.)
J. Giménez and L. Màrquez. Linguistic measures for automatic machine translation evaluation. *Machine Translation*, 24(3-4):209–240, 2010. (Cited on page 85.)
A. Gliozzo and C. Strapparava. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 553–560, 2006. (Cited on page 46.)
L. Guillou, C. Hardmeier, P. Nakov, S. Stymne, J. Tiedemann, Y. Versley, M. Cettolo, B. Webber, and A. Popescu-Belis. Findings of the 2016 wmt shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation*, pages 525–542, 2016. (Cited on page 100.)
L. K. Guillou. *Incorporating pronoun function into statistical machine translation*. PhD thesis, The University of Edinburgh, 2016. (Cited on page 100.)
N. Habash and O. Rambow. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 573–580, 2005. (Cited on pages 37 and 86.)
B. Haddow and P. Koehn. Analysing the effect of out-of-domain data on SMT systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 422–432, 2012. (Cited on pages 33, 85, and 137.)
M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009. (Cited on page 57.)
C. Hardmeier. *Discourse in statistical machine translation*. PhD thesis, Acta Universitatis Upsaliensis, 2014. (Cited on page 100.)
C. Hardmeier, J. Tiedemann, and J. Nivre. Latent anaphora resolution for cross-lingual pronoun prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 380–391, 2013. (Cited on page 100.)
C. Hardmeier, P. Nakov, S. Stymne, J. Tiedemann, Y. Versley, and M. Cettolo. Pronoun-focused mt and cross-lingual pronoun prediction: Findings of the 2015 discomt shared task on pronoun translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16, 2015. (Cited on page 100.)
E. Hasler. *Dynamic topic adaptation for improved contextual modelling in statistical machine translation*. PhD thesis, University of Edinburgh, 2014. (Cited on page 34.)
E. Hasler, B. Haddow, and P. Koehn. Sparse lexicalised features and topic adaptation for smt. In *Proceedings of the 9th International Workshop on Spoken Language Translation*, pages 268–275, 2012. (Cited on pages 35 and 70.)
E. Hasler, P. Blunsom, P. Koehn, and B. Haddow. Dynamic topic adaptation for phrase-based mt. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 328–337, 2014a. (Cited on pages 35 and 70.)
E. Hasler, B. Haddow, and P. Koehn. Combining domain and topic adaptation for SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 139–151, 2014b. (Cited on pages 35, 70, and 136.)
E. Hasler, B. Haddow, and P. Koehn. Dynamic topic adaptation for SMT using distributional profiles. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 445–456, 2014c. (Cited on pages 35 and 70.)
E. Hasler, A. de Gispert, F. Stahlberg, A. Waite, and B. Byrne. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 2016. (Cited on page 40.)
S. Hewavitharana, D. Mehay, S. Ananthakrishnan, and P. Natarajan. Incremental topic-based translation model adaptation for conversational spoken language translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 697–701, 2013. (Cited on pages 35 and 70.)
A. S. Hildebrand, M. Eck, S. Vogel, and A. Waibel. Adaptation of the translation model for statistical machine translation based on information retrieval. In *Proceedings of the Conference of the European Association for Machine Translation*, pages 133–142, 2005. (Cited on page 31.)
S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. (Cited on pages 19 and 119.)
C. Hokamp and Q. Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 1535–1546, 2017. (Cited on page 117.)

J. Holmes and M. Meyerhoff. *The handbook of language and gender*, volume 25. John Wiley & Sons, 2008. (Cited on page 109.)

M. Hopkins and J. May. Tuning as ranking. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362, 2011. (Cited on pages 16 and 24.)

A. Irvine and C. Callison-Burch. Hallucinating phrase translations for low resource MT. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 160–170, 2014. (Cited on page 40.)

A. Irvine, J. Morgan, M. Carpuat, H. Daumé III, and D. S. Munteanu. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440, 2013a. (Cited on pages 33, 39, 70, 85, 92, and 93.)

A. Irvine, C. Quirk, and H. Daumé III. Monolingual marginal matching for translation model adaptation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1088, 2013b. (Cited on page 32.)

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, et al. The ICSI meeting corpus. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I–364. IEEE, 2003. (Cited on page 100.)

L. Jehl, F. Hieber, and S. Riezler. Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 410–421, 2012. (Cited on pages 45 and 85.)

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142, 1998. (Cited on page 57.)

G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995. (Cited on page 57.)

N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013. (Cited on page 18.)

N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099*, 2016. (Cited on page 118.)

J. Karlgren. The wheres and whyfores for studying text genre computationally. In *Workshop on Style and Meaning in Language, Art, Music, and Design*, 2004. (Cited on page 34.)

J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th International conference on Computational Linguistics*, pages 1071–1075, 1994. (Cited on page 45.)

B. Kessler, G. Numberg, and H. Schütze. Automatic detection of text genre. In *Proceedings of the eighth conference of the European chapter of the Association for Computational Linguistics (EACL)*, pages 32–38, 1997. (Cited on page 45.)

Y. Kim and A. M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, 2016. (Cited on page 118.)

R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184, 1995. (Cited on page 24.)

K. Knight. Decoding complexity in word-replacement translation models. *Computational linguistics*, 25(4): 607–615, 1999. (Cited on page 16.)

C. Kobus, J. Crego, and J. Senellart. Domain control for neural machine translation. *arXiv preprint arXiv:1612.06140*, 2016. (Cited on page 117.)

P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pages 79–86, 2005. (Cited on pages 1, 12, 47, and 51.)

P. Koehn. *Statistical machine translation*. Cambridge University Press, 2009. (Cited on pages 12 and 18.)

P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, 2017. (Cited on pages 21, 116, 119, and 137.)

P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, 2007. (Cited on pages 1, 32, and 137.)

P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, 2003. (Cited on pages 14, 24, and 89.)

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International*

Workshop on Spoken Language Translation, 2005. (Cited on page 13.)

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, 2007. (Cited on pages 24 and 91.)

J. Kreutzer, A. Sokolov, and S. Riezler. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1503–1513, 2017. (Cited on page 117.)

T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998. (Cited on page 33.)

P. Langlais. Improving a general-purpose statistical translation engine by terminological lexicons. In *COMPUTERM 2002: second international workshop on computational terminology*, pages 1–7, 2002. (Cited on page 30.)

M. M. Lauzen. The celluloid ceiling: Behind-the-scenes employment of women on the top 100, 250, and 500 films of 2015. Technical report, Center for the study of women in television and film, 2016. (Cited on page 109.)

D. Y. Lee. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle. *Language Learning & Technology*, 5(3):37–72, September 2001. (Cited on pages 34 and 97.)

Y.-B. Lee and S. H. Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150, 2002. (Cited on pages 29, 33, 45, and 67.)

O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, 2014. (Cited on page 18.)

W. D. Lewis and S. Eetemadi. Dramatically reducing training data size through vocabulary saturation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 281–291, 2013. (Cited on pages 117 and 120.)

J. J. Li, M. Carpuat, and A. Nenkova. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 283–288, 2014. (Cited on page 100.)

W. Ling, C. Dyer, A. W. Black, and I. Trancoso. Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84, 2013a. (Cited on pages 85 and 96.)

W. Ling, G. Xiang, C. Dyer, A. Black, and I. Trancoso. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 176–186, 2013b. (Cited on pages 45 and 85.)

P. Lison and R. Meena. Automatic turn segmentation for movie & tv subtitles. In *Proceedings of the Spoken Language Technology Workshop (SLT)*, pages 245–252, 2016. (Cited on page 99.)

P. Lison and J. Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016. (Cited on pages 99 and 122.)

M.-T. Luong and C. D. Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 76–79, 2015. (Cited on pages 117 and 121.)

M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015a. (Cited on pages 20, 21, 24, and 25.)

M.-T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Learning (ACL-CoNLL)*, pages 11–19, 2015b. (Cited on pages 115 and 119.)

X. Ma. Champollion: A robust parallel text sentence aligner. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*, pages 489–492, 2006. (Cited on pages 49 and 100.)

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947. (Cited on page 73.)

S. Matsoukas, A.-V. I. Rosti, and B. Zhang. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages

708–717, 2009. (Cited on pages 1, 33, 39, 68, and 94.)

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, et al. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, 2005. (Cited on page 100.)

T. Meyer and B. Webber. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, 2013. (Cited on page 100.)

T. Meyer, A. Popescu-Belis, N. Hajlaoui, and A. Gesmundo. Machine translation of labeled discourse connectives. In *Proceedings of the Tenth Biennial Conference of the Association for Machine Translation in the Americas*, pages 129–138, 2012. (Cited on page 100.)

T. Mikolov. Recurrent neural network based language model. In *INTERSPEECH*, page 3, 2010. (Cited on pages 19 and 119.)

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a. (Cited on page 19.)

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b. (Cited on page 18.)

S. Mirkin and L. Besacier. Data selection for compact adapted SMT models. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 301–314, 2014. (Cited on page 117.)

S. Mirkin and J.-L. Meunier. Personalized machine translation: Predicting translational preferences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, 2015. (Cited on page 108.)

S. Mirkin, S. Nowson, C. Brun, and J. Perez. Motivating personality-aware machine translation. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, 2015. (Cited on page 108.)

R. C. Moore. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 135–144, 2002. (Cited on page 49.)

R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 220–224, 2010. (Cited on pages 31, 115, 117, and 119.)

A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *Proceedings of the 26th European Conference on Information Retrieval*, pages 181–196, 2004. (Cited on page 56.)

D. S. Munteanu and D. Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005. (Cited on pages 32 and 45.)

D. S. Munteanu and D. Marcu. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 81–88, 2006. (Cited on page 45.)

G. Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017. (Cited on pages 19 and 24.)

S. Nießen, S. Vogel, H. Ney, and C. Tillmann. A dp based search algorithm for statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 960–967, 1998. (Cited on page 17.)

E. W. Noreen. *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience, 1989. (Cited on page 105.)

F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, 2003. (Cited on page 16.)

F. J. Och and H. Ney. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, 2002. (Cited on page 16.)

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. (Cited on pages 24 and 89.)

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. (Cited on pages 25 and 83.)

P. Pecina, A. Toral, A. Way, V. Papavassiliou, P. Prokopidis, and M. Giagkou. Towards using web-crawled data for domain adaptation in statistical machine translation. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 297–304, 2011. (Cited on pages 32 and 45.)

P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, J. Van Genabith, and R. Athena. Domain adaptation of statistical machine translation using web-crawled resources: a case study. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 145–152, 2012. (Cited on page 45.)

P. Pecina, A. Toral, V. Papavassiliou, P. Prokopidis, A. Tamchyna, A. Way, and J. van Genabith. Domain adaptation of statistical machine translation with domain-focused web crawling. *Language resources and evaluation*, 49(1):147–193, 2015. (Cited on pages 32 and 45.)

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. (Cited on page 19.)

P. Petrenz. Cross-lingual genre classification. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 11–21, 2012. (Cited on page 46.)

P. Petrenz and B. Webber. Robust cross-lingual genre classification through comparable corpora. In *The 5th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 1–9, 2012. (Cited on pages 46 and 73.)

V. Petukhova, R. Agerri, M. Fishel, S. Penkale, A. del Pozo, M. S. Maucec, A. Way, P. Georgakopoulou, and M. Volk. Sumat: Data collection and parallel corpus compilation for machine translation of subtitles. In *LREC*, pages 21–28, 2012. (Cited on page 99.)

B. Plank, A. Johannsen, and A. Søgaard. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 968–973, 2014. (Cited on page 40.)

J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*. MIT Press, 1998. (Cited on page 57.)

M. Popović and H. Ney. Error analysis of verb inflections in spanish translation output. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 99–103, 2006. (Cited on page 85.)

M. Popović and H. Ney. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688, 2011. (Cited on page 85.)

X. Qi and B. D. Davison. Web page classification: Features and algorithms. *ACM Computing Surveys*, 41(2): 12:1–12:31, 2009. (Cited on page 44.)

R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993. (Cited on page 57.)

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010. (Cited on page 74.)

S. Riezler and J. T. Maxwell. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, 2005. (Cited on pages 25, 63, and 110.)

J. Roturier and A. Bensadoun. Evaluation of MT systems to translate user generated content. In *Proceedings of the XIII Machine Translation Summit*, pages 244–251, 2011. (Cited on page 85.)

S. Ruder and B. Plank. Learning to select data for transfer learning with bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. (Cited on page 120.)

N. Ruiz and M. Federico. Topic adaptation for lecture translation through bilingual latent semantic models. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 294–302, 2011. (Cited on page 35.)

N. Ruiz and M. Federico. MDI adaptation for the lazy: avoiding normalization in LM adaptation for lecture translation. In *Proceedings of the 9th International Workshop on Spoken Language Translation*, pages 244–251, 2012. (Cited on page 35.)

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. (Cited on page 22.)

M. Santini. State-of-the-art on automatic genre identification. Technical Report ITRI-04-03, Information Technology Research Institute, University of Brighton, 2004. (Cited on page 34.)

D. Schlangen. Modelling dialogue: Challenges and approaches. *Künstliche Intelligenz*, 3:23–28, 2005. (Cited on page 97.)

H. Schwenk. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the 5th International Workshop on Spoken Language Translation*, pages 182–189, 2008. (Cited on page 32.)

F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1): 1–47, 2002. (Cited on pages 45 and 56.)

R. Sennrich. Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 185–192, 2012a. (Cited on page 32.)

R. Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 539–549, 2012b. (Cited on pages 1 and 32.)

R. Sennrich, H. Schwenk, and W. Aransa. A multi-domain translation model framework for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 832–840, 2013. (Cited on page 32.)

R. Sennrich, B. Haddow, and A. Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, 2016a. (Cited on page 137.)

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, 2016b. (Cited on pages 12, 117, and 121.)

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016c. (Cited on pages 119 and 122.)

R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. Miceli Barone, J. Mokry, and M. Nadejde. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, 2017. (Cited on page 23.)

S. Sharoff. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of the 3rd Web as Corpus Workshop*, 2007. (Cited on page 46.)

S. Sharoff, Z. Wu, and K. Markert. The web library of babel: evaluating genre collections. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 3063–3070, 2010. (Cited on page 46.)

X. Shi, I. Padhi, and K. Knight. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, 2016. (Cited on page 20.)

K. Shinoda. Speaker adaptation techniques for automatic speech recognition. In *Proceedings of APSIPA ASC*, 2011. (Cited on page 108.)

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas*, pages 223–231, 2006. (Cited on page 83.)

L. Song, H. Mi, Y. Lü, and Q. Liu. Bagging-based system combination for domain adaptation. In *Proceedings of the XIII Machine Translation Summit*, pages 293–298, 2011. (Cited on page 32.)

M. Spousta, M. Marek, and P. Pecina. Victor: the web-page cleaning tool. In *4th Web as Corpus Workshop: Can we beat Google?*, pages 12–17, 2008. (Cited on page 44.)

N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, 15(1):1929–1958, 2014. (Cited on page 25.)

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics*, pages 808–814, 2000. (Cited on page 45.)

B. Stein and S. Meyer Zu Eissen. Distinguishing topic from genre. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW)*, pages 449–456, 2006. (Cited on pages 29, 33, and 67.)

A. Stolcke et al. Srilm-an extensible language modeling toolkit. In *INTERSPEECH*, 2002. (Cited on pages 24 and 119.)

J. Su, D. Xiong, Y. Liu, X. Han, H. Lin, J. Yao, and M. Zhang. A context-aware topic model for statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 229–238, 2015. (Cited on pages 35 and 70.)

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014. (Cited on pages 18, 19, 20, and 115.)

J. M. Swales. *Genre Analysis*. Cambridge University Press, Cambridge, UK, 1990. (Cited on page 34.)

Y.-C. Tam, I. Lane, and T. Schultz. Bilingual lsa-based adaptation for statistical machine translation. *Machine translation*, 21(4):187–207, 2007. (Cited on pages 35 and 70.)

D. Tannen. *Gender and discourse*. Oxford University Press, 1994. (Cited on page 109.)
J. Tiedemann. Synchronizing translated movie subtitles. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1902–1906, 2008. (Cited on page 99.)
J. Tiedemann. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, pages 237–248, 2009a. (Cited on pages 99 and 122.)
J. Tiedemann. Translating questions for cross-lingual QA. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages 112–119, 2009b. (Cited on pages 100 and 106.)
J. Tiedemann. Finding alternative translations in a large corpus of movie subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016. (Cited on pages 99 and 101.)
C. Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 101–104, 2004. (Cited on pages 14 and 24.)
C. Tillmann. A beam-search extraction algorithm for comparable data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 225–228, 2009. (Cited on page 45.)
C. Tillmann and J.-m. Xu. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 93–96, 2009. (Cited on page 45.)
C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga. A DP-based search using monotone alignments in statistical translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 289–296, 1997. (Cited on page 17.)
K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, 2003. (Cited on page 57.)
H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter. In *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, pages 168–171, 2005. (Cited on pages 50 and 86.)
F. Ture and E. Boschee. Learning to translate for multilingual question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 573–584, 2016. (Cited on pages 100 and 106.)
N. Ueffing. Using monolingual source-language data to improve mt performance. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 174–181, 2006. (Cited on page 32.)
M. van der Wees, A. Bisazza, and C. Monz. Five shades of noise: Analyzing machine translation errors in user-generated text. In *Proceedings of the First Workshop on Noisy User-generated Text (WNUT)*, pages 28–37, 2015a. (Cited on page 9.)
M. van der Wees, A. Bisazza, and C. Monz. Translation model adaptation using genre-revealing text features. In *Proceedings of the Second Workshop on Discourse in Machine Translation (DiscoMT)*, pages 132–141, 2015b. (Cited on page 9.)
M. van der Wees, A. Bisazza, W. Weerkamp, and C. Monz. What's in a domain? Analyzing genre and topic differences in statistical machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Learning (ACL-CoNLL)*, pages 560–566, 2015c. (Cited on page 9.)
M. van der Wees, A. Bisazza, and C. Monz. Measuring the effect of conversational aspects on machine translation quality. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2571–2581, 2016a. (Cited on page 9.)
M. van der Wees, A. Bisazza, and C. Monz. A simple but effective approach to improve arabizi-to-english statistical machine translation. In *Proceedings of the Second Workshop on Noisy User-generated Text (WNUT)*, pages 100–107, 2016b. (Cited on page 10.)
M. van der Wees, A. Bisazza, and C. Monz. Dynamic data selection for neural machine translation. In *Proceedings of EMNLP2017*, pages 1411–1421, 2017. (Cited on page 10.)
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. (Cited on page 21.)
D. Vilar, J. Xu, L. F. dHaro, and H. Ney. Error analysis of statistical machine translation output. In *Proceedings of LREC 2006: Fifth International Conference on Language Resources and Evaluation*, pages 489–492, 2006. (Cited on page 85.)
S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational Linguistics*, pages 836–841. Association for Computational

Linguistics, 1996. (Cited on page 13.)
M. Volk. The automatic translation of film subtitles. a machine translation success story? *JLCL*, 24(3): 115–128, 2009. (Cited on page 99.)
M. Volk, R. Sennrich, C. Hardmeier, and F. Tidström. Machine translation of tv subtitles for large scale production. In *Proceedings of the JEC'10*, pages 53–62, 2010. (Cited on page 99.)
M. A. Walker, G. I. Lin, and J. Sawyer. An annotated corpus of film dialogue for learning and characterizing character style. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1373–1378, 2012. (Cited on pages 99 and 100.)
L. Wang, X. Zhang, Z. Tuy, A. Way, and Q. Liu. Automatic construction of discourse corpora for dialogue translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016. (Cited on pages 99 and 100.)
R. Wang, A. Finch, M. Utiyama, and E. Sumita. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 560–566, 2017a. (Cited on page 118.)
R. Wang, M. Utiyama, L. Liu, K. Chen, and E. Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017b. (Cited on page 118.)
W. Wang, K. Macherey, W. Macherey, F. Och, and P. Xu. Improved domain adaptation for statistical machine translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, 2012. (Cited on pages 32, 46, and 137.)
Y.-Y. Wang and A. Waibel. Decoding algorithm in statistical machine translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 366–372, 1997. (Cited on page 17.)
S. Whittaker. Theories and methods in mediated communication. In *The Handbook of Discourse Processes*, pages 243–286. Erlbaum, 2003. (Cited on page 98.)
R. Wodak. *Gender and discourse*. Sage, 1997. (Cited on page 109.)
H. Wu and H. Wang. Improving domain-specific word alignment with a general bilingual corpus. In *Conference of the Association for Machine Translation in the Americas*, pages 262–271, 2004. (Cited on page 33.)
H. Wu, H. Wang, and Z. Liu. Alignment model adaptation for domain-specific word alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 467–474, 2005. (Cited on page 33.)
Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. (Cited on page 118.)
J. Xu, Y. Deng, Y. Gao, and H. Ney. Domain dependent statistical machine translation. In *Proceedings of the XI Machine Translation Summit*, pages 515–520, 2007. (Cited on pages 32 and 46.)
H. Yamamoto and E. Sumita. Bilingual cluster based models for statistical machine translation. *IEICE TRANSACTIONS on Information and Systems*, 91(3):588–597, 2008. (Cited on page 32.)
K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumit. Method of selecting training data to build a compact and efficient translation model. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 655–660, 2008. (Cited on pages 31 and 117.)
S. Yitzhaki. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron*, 61(2):285–316, 2003. (Cited on page 106.)
F. Yvon. Rewriting the orthography of SMS messages. *Natural Language Engineering*, 16(2):133–159, 2010. (Cited on page 96.)
R. Zens, F. J. Och, and H. Ney. Phrase-based statistical machine translation. In *Proceedings of the German Conference on Artificial Intelligence (KI)*, 2002. (Cited on page 17.)
D. Zhang, J. Kim, J. Crego, and J. Senellart. Boosting neural machine translation. *arXiv preprint arXiv:1612.06138*, 2016. (Cited on page 118.)
A. Zollmann and A. Venugopal. Syntax augmented machine translation via chart parsing. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, 2006. (Cited on page 11.)
B. Zoph, D. Yuret, J. May, and K. Knight. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*, 2016. (Cited on pages 116, 117, 119, and 121.)

Summary

Machine translation (MT) uses software to translate texts written in one language (for example German) to another language (for example English). Modern-day MT systems are built using large amounts of example translations between the two languages of interest, so-called *parallel corpora*. If parallel corpora are sufficiently large and of high quality, created by professional translators who adhere to language standardizations, an MT system can produce good translations.

However, sizable high-quality parallel corpora only exist for a limited number of translation tasks, or *domains*, such as parliamentary proceedings. The picture looks less bright when training data is scarce for a domain of interest, for example if one wishes to translate medical texts. In such cases, the available training data differs from the translation task in both writing style and vocabulary, a mismatch that can cause large drops in translation quality. In recent years, this problem has been addressed by *domain adaptation*, in which an MT system is adapted to the domain of interest and translation quality is often improved.

Unfortunately, the notion of a domain is not uniformly defined. Typically, domain means 'different data set,' and is thus a hard-labeled concept that is often directly used to optimize an MT system. This definition ignores three important facts: first, documents or sentences *within* a single domain may vary at many levels, such as *topics*, *genres*, and *register*, which may be useful information to adapt an MT system. Second, some domains or genres may require different strategies than others to improve translation quality due to their inherent differences. Third, available domain labels may not provide the most useful information for effective adaptation.

To shed light on the concept of a domain and its impact on MT, the core question in this thesis is: *"What's in a domain?"* Guided by this question, we distinguish various aspects that together make up a domain, i.e., topic, genre, register, dialogue acts, speakers, and speaker gender. We study to what extent MT output differs among these aspects, and how we can use them to perform fine-grained adaptation for MT. We are particularly interested in *informal* and *conversational* genres, which lack standardization and are notorious for poor MT output. In addition, we aim to develop methods that do not, or at most partially, rely on manual domain or genre information.

By studying what's in a domain and showing how we can use different aspects of language to improve MT, we take in this thesis a step forward towards fine-grained adaptation for machine translation.

Samenvatting

Computervertaalsystemen (Engels: machine translation (MT) systems) gebruiken software om teksten geschreven in één taal (bijvoorbeeld Nederlands) te vertalen naar een andere taal (bijvoorbeeld Engels). Hedendaagse computervertaalsystemen worden gebouwd door te leren van grote hoeveelheden voorbeeldvertalingen tussen twee talen, zogeheten *parallelle corpora*. Als zulke parallelle corpora voldoende groot zijn en van goede kwaliteit, vaak vertaald door professionele vertalers, kan een computervertaalsysteem vertalingen van goede kwaliteit genereren.

Parallelle corpora van goede kwaliteit zijn echter enkel beschikbaar voor een beperkt aantal *domeinen*, zoals nieuws of parlementaire verslaggevingen. De situatie is minder rooskleurig voor domeinen waarvoor meertalige *training data* schaars is, wat bijvoorbeeld het geval is voor medische teksten of berichten op sociale media. In zulke gevallen zorgen de verschillen in schrijfstijl en vocabulaire tussen de voorbeeldvertalingen en de te vertalen tekst ervoor dat de vertaalkwaliteit drastisch verslechtert. In de afgelopen jaren is er veel aandacht besteed om dit probleem op te lossen door middel van domeinadaptatie (Engels: domain adaptation). In dit proces wordt het vertaalsysteem aangepast aan het domein van belang waardoor de vertaalkwaliteit voor dit domein verbetert.

Helaas is het concept *domein* niet eenduidig gedefinieerd in de huidige literatuur. Meestal betekent een nieuw domein een 'andere dataset' en wordt informatie over de oorsprong van deze dataset direct gebruikt om een vertaalsysteem te optimaliseren. Deze definitie negeert drie belangrijke feiten: ten eerste, documenten of zinnen binnen een domein kunnen variëren op vele verschillende niveaus, zoals qua *onderwerp*, *tekst-genre* of *taalgebruik*. Variatie op deze niveaus kan belangrijke informatie bevatten om een vertaalsysteem aan te passen. Ten tweede, sommige domeinen of tekst-genres vereisen specifieke strategieën om vertaalkwaliteit te verbeteren dankzij inherente kenmerken van die domeinen. Ten derde, beschikbare domein-labels bevatten niet per definitie de meest waardevolle informatie voor effectieve adaptatie.

Om inzicht te krijgen in het concept domein en de impact van domeinen op computervertaalsystemen, stelt dit proefschrift de vraag: "Wat omvat een domein?" Aan de hand van deze vraag onderscheiden we verscheidene aspecten die tezamen een domein definiëren, zoals onderwerp, tekst-genre, taalgebruik, dialoogkenmerken, sprekers en het geslacht van sprekers. We bestuderen in hoeverre vertaalkwaliteit varieert binnen elk van deze aspecten, en hoe we deze aspecten kunnen gebruiken om adaptatie van vertaalsystemen op verschillende niveaus te bewerkstelligen. Hierbij zijn we specifiek geïnteresseerd in *informele* teksten en *conversaties*, die beide worden gekenmerkt door een gebrek aan standaardisatie en een notoir slechte vertaalkwaliteit. Daarnaast beogen we methodes te ontwikkelen die niet, of slechts gedeeltelijk, afhankelijk zijn van handmatig gedefiniëerde domein-informatie.

Door te bestuderen wat een domein omvat en aan te tonen hoe we verschillende aspecten van taal kunnen benutten om computervertaalsystemen te verbeteren, nemen we in dit proefschrift een belangrijke stap richting verbeterde adaptatie voor computervertaalsystemen.

Machine translation (MT) uses software to translate texts written in one language (for example German) to another language (for example English). Modern-day MT systems are built using large amounts of example translations between the two languages of interest, so-called parallel corpora. If parallel corpora are sufficiently large and of high quality, created by professional translators who adhere to language standardizations, an MT system can produce good translations.

However, sizable high-quality parallel corpora only exist for a limited number of translation tasks, or domains, such as parliamentary proceedings. The picture looks less bright when training data is scarce for a domain of interest, for example if one wishes to translate medical texts. In such cases, the available training data differs from the translation task in both writing style and vocabulary, a mismatch that can cause large drops in translation quality. In recent years, this problem has been addressed by domain adaptation, in which an MT system is adapted to the domain of interest and translation quality is often improved.

Unfortunately, the notion of a domain is not uniformly defined. Typically, domain means 'different data set,' and is thus a hard-labeled concept that is often directly used to optimize an MT system. This definition ignores three important facts: first, documents or sentences within a single domain may vary at many levels, such as topics, genres, and register, which may be useful information to adapt an MT system. Second, some domains or genres may require different strategies than others to improve translation quality due to their inherent differences. Third, available domain labels may not provide the most useful information for effective adaptation.

To shed light on the concept of a domain and its impact on MT, the core question in this thesis is: "What's in a domain?" Guided by this question, we distinguish various aspects that together make up a domain, i.e., topic, genre, register, dialogue acts, speakers, and speaker gender. We study to what extent MT output differs among these aspects, and how we can use them to perform fine-grained adaptation for MT. We are particularly interested in informal and conversational genres, which lack standardization and are notorious for poor MT output. In addition, we aim to develop methods that do not, or at most partially, rely on manual domain or genre information.

By studying what's in a domain and showing how we can use different aspects of language to improve MT, we take in this thesis a step forward towards fine-grained adaptation for machine translation.

ISBN 978-94-6182-852-1
9 789461 828521