Evaluation of Machine Translation Performance Across Multiple Genres and Languages

van der Wees, M.; Bisazza, A.; Monz, C.

Publication date
2018

Document Version
Final published version

Published in
LREC 2018: Eleventh International Conference on Language Resources and Evaluation

License
CC BY-NC

Citation for published version (APA):

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, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Evaluation of Machine Translation Performance Across Multiple Genres and Languages

Marlies van der Wees¹, Arianna Bisazza², Christof Monz¹
¹Informatics Institute, University of Amsterdam
²Leiden Institute of Advanced Computer Science, Leiden University

Abstract

In this paper, we present evaluation corpora covering four genres for four language pairs that we harvested from the web in an automated fashion. We use these multi-genre benchmarks to evaluate the impact of genre differences on machine translation (MT). We observe that BLEU score differences between genres can be large and that, for all genres and all language pairs, translation quality improves when using four genre-optimized systems rather than a single genre-agnostic system. Finally, we train and use genre classifiers to route test documents to the most appropriate genre systems. The results of these experiments show that our multi-genre benchmarks can serve to advance research on text genre adaptation for MT.

Keywords: Machine translation, parallel benchmarks, text genres, genre adaptation

1. Introduction

Text genre differences have shown to affect the output quality of statistical machine translation (SMT) systems: SMT systems trained on one genre often achieve poor performance when used for translating another genre (Foster and Kuhn, 2007; Matsoukas et al., 2009; Wang et al., 2012, among others). In addition, even if different genres in a test set are both present in equal amounts in the bilingual training data, performance differences between the test genres can be large, mostly due to poor model coverage for certain genres (van der Wees et al., 2015a; van der Wees et al., 2015b).

In this paper, we evaluate the impact of genre differences on phrase-based SMT for a diverse set of language pairs, covering both commonly and rarely studied language pairs. For common language pairs, parallel training data 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. To alleviate this problem, we present in this paper novel parallel training and evaluation corpora covering four genres for four language pairs that we automatically harvested from the web.

Next, we evaluate the usefulness of the newly collected bilingual resources by exploiting them for genre adaptation of SMT systems. Most existing adaptation approaches depend on the availability of provenance information and make the strong assumption that a translation task has known domain, genre or topic that is exploited to adapt the system (Matsoukas et al., 2009; Foster et al., 2010; Bisazza and Federico, 2012; Chen et al., 2013; Chen et al., 2014; Kobus et al., 2016; Sennrich et al., 2016; Freitag and Al-Onaizan, 2016; Chu et al., 2017, among others). While this is a fair assumption in a controlled research setting, it is less realistic in real world applications, such as general-purpose online MT services. In this paper, we provide the SMT system with a test document of unknown origin, and we show that we can use automatic genre classification to guide each test document to the most appropriate pre-trained system. While similar setups have been used in previous work (Xu et al., 2007; Banerjee et al., 2010; Pecina et al., 2011; Wang et al., 2012; Pecina et al., 2015), we are the first to extend this setup to four genres and four language pairs. Finally, we show that an adaptation method based on automatic classifiers also improves translation quality for genres with no parallel training data available.

2. Multi-genre benchmarks

In this section, we describe the construction of multi-genre corpora for four language pairs and four genres, which we obtained using an automated web-harvesting process.

2.1. Language pairs and genres

While most research in MT is evaluated on a small number of well resourced language pairs and domains or genres, 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 paper on the following language pairs: Arabic→English, Chinese→English, Bulgarian→English, and Persian→English. For each of these language pairs we consider four different genres: news, as it can be found in (online) newspapers and in transcripts of broadcast news; editorial, covering Op-Ed pieces in (online) newspapers, that represent a subjective, and unlike news less matter-of-fact point of view; colloquial, covering informal conversation such as blog comments and Internet forum discussions; and speech, covering speeches for which transcripts are available such as TED talks and other public speeches.

2.2. Benchmark construction

For the language pairs and genres of interest, we collect parallel corpora from the web from twenty different websites, each covering at least one of our genres of interest. All websites contain manual translations at the sentence level.

¹Sixteen websites contain news documents, six websites contain editorial documents, eight websites contain colloquial documents, and three websites contain speech transcripts.
Table 1: Specifications of the harvested multi-genre training, development and test sets for four language pairs. Tokens are counted on the English side. We make the evaluation corpora available at [http://ilps.science.uva.nl/resources/genre-benchmarks](http://ilps.science.uva.nl/resources/genre-benchmarks).

<table>
<thead>
<tr>
<th>Genre</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquial</td>
<td>273K</td>
<td>8.9M</td>
<td>1.5K</td>
<td>77.3K</td>
<td>1.5K</td>
<td>73.0K</td>
</tr>
<tr>
<td>Editorial</td>
<td>156K</td>
<td>4.7M</td>
<td>1.5K</td>
<td>45.6K</td>
<td>1.5K</td>
<td>47.3K</td>
</tr>
<tr>
<td>News</td>
<td>600K</td>
<td>18.0M</td>
<td>1.5K</td>
<td>50.4K</td>
<td>1.5K</td>
<td>48.1K</td>
</tr>
<tr>
<td>Speech</td>
<td>140K</td>
<td>3.4M</td>
<td>1.5K</td>
<td>35.7K</td>
<td>1.5K</td>
<td>38.7K</td>
</tr>
<tr>
<td>Total</td>
<td>1.2M</td>
<td>35.0M</td>
<td>6.0K</td>
<td>209K</td>
<td>6.0K</td>
<td>207K</td>
</tr>
</tbody>
</table>

(a) Arabic→English data.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquial</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1.4K</td>
<td>33.9K</td>
</tr>
<tr>
<td>Editorial</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>178</td>
<td>5.1K</td>
</tr>
<tr>
<td>News</td>
<td>215K</td>
<td>5.3M</td>
<td>1.2K</td>
<td>30.2K</td>
<td>2.0K</td>
<td>49.5K</td>
</tr>
<tr>
<td>Speech</td>
<td>206K</td>
<td>3.9M</td>
<td>1.2K</td>
<td>22.5K</td>
<td>2.0K</td>
<td>44.6K</td>
</tr>
<tr>
<td>Total</td>
<td>422K</td>
<td>9.2M</td>
<td>2.4K</td>
<td>52.7K</td>
<td>5.6K</td>
<td>133K</td>
</tr>
</tbody>
</table>

(b) Chinese→English data.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquial</td>
<td>629K</td>
<td>16.4M</td>
<td>1.5K</td>
<td>40.3K</td>
<td>1.5K</td>
<td>37.7K</td>
</tr>
<tr>
<td>Editorial</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>600</td>
<td>19.4K</td>
</tr>
<tr>
<td>News</td>
<td>618K</td>
<td>16.8M</td>
<td>1.5K</td>
<td>44.5K</td>
<td>1.5K</td>
<td>47.4K</td>
</tr>
<tr>
<td>Speech</td>
<td>119K</td>
<td>2.5M</td>
<td>1.5K</td>
<td>31.2K</td>
<td>1.5K</td>
<td>35.6K</td>
</tr>
<tr>
<td>Total</td>
<td>1.4M</td>
<td>35.7M</td>
<td>4.5K</td>
<td>116K</td>
<td>5.1K</td>
<td>140K</td>
</tr>
</tbody>
</table>

(c) Bulgarian→English data.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
<th>Lines</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquial</td>
<td>55K</td>
<td>1.7M</td>
<td>1.5K</td>
<td>42.5K</td>
<td>1.4K</td>
<td>35.8K</td>
</tr>
<tr>
<td>Editorial</td>
<td>370K</td>
<td>10.2M</td>
<td>1.5K</td>
<td>43.1K</td>
<td>1.5K</td>
<td>42.6K</td>
</tr>
<tr>
<td>News</td>
<td>584K</td>
<td>16.4M</td>
<td>1.5K</td>
<td>39.2K</td>
<td>1.5K</td>
<td>35.8K</td>
</tr>
<tr>
<td>Speech</td>
<td>146K</td>
<td>3.3M</td>
<td>1.5K</td>
<td>42.6K</td>
<td>1.5K</td>
<td>37.5K</td>
</tr>
<tr>
<td>Total</td>
<td>1.2M</td>
<td>31.6M</td>
<td>6.0K</td>
<td>169K</td>
<td>5.9K</td>
<td>152K</td>
</tr>
</tbody>
</table>

(d) Persian→English data.

Table 2: English example sentences for four genres in the web-harvested evaluation corpora.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Example sentence(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colloquial</td>
<td>Ministers should be sitting and attending the oath, like in Italy.</td>
</tr>
<tr>
<td>Editorial</td>
<td>This may sound like pie in the sky, but we have already tasted it in Africa, where Sierra Leone’s agenda for prosperity 2013–2017 and the Liberia Vision 2030 exemplify the potential of such programs.</td>
</tr>
<tr>
<td>News</td>
<td>She is not only the first Saudi woman to ever attempt the climb but also the youngest Arab to make it to the top of the world’s highest peak.</td>
</tr>
<tr>
<td>Speech</td>
<td>These are just a few of the milestones of recent progress. I have another reason to be optimistic. I know global health is guided by the right values.</td>
</tr>
</tbody>
</table>

3. Evaluating genre differences in SMT

In this section, we use our newly assembled resources to evaluate SMT performance across different genres and language pairs.
3.1. Experimental setup
All SMT systems in this paper are trained using an in-house phrase-based SMT system similar to Moses (Koehn et al., 2007). To train our systems, we use our web-crawled corpora, supplemented with commonly used training data, if available: LDC corpora for Arabic→English and Chinese→English, and Europarl data (Koehn, 2005) for Bulgarian→English. In addition, we use a 5-gram language model that linearly interpolates various Gigaword subcorpora with the English sides of the bilingual training corpora. To evaluate the effect of our new bilingual resources, we do not vary the language model between experiments.

In order to create genre-specific SMT systems, we have to adequately use the available data. Simply concatenating the different corpora yields a general SMT 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 the fill-up technique proposed by 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.

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 mixed 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.

3.2. Results
Tables 3a–3d show the translation quality results for all language pairs. For each language pair, we measure case-insensitive BLEU (Papineni et al., 2002) 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 a single reference translation. 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.

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. For example, editorial Persian→English data is scarce, so for Persian editorial documents we have to resort to our colloquial, news, speech or mixed 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.

4. Genre adaptation using automatic classifiers
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 MT system. Adapting an MT system using this strategy involves two steps: training accurate
genre classifiers (§4.1) and incorporating these classifiers into an end-to-end MT pipeline (§4.2).

Adaptation using automatic classifiers has been applied in a few previous efforts [Xu et al., 2007; Banerjee et al., 2010; Wang et al., 2012]. However, all three methods limit their application to one language pair covering 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.

4.1. Training genre classifiers

Since we apply our genre classifiers to different languages, 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. To train the classifiers we randomly select documents from the training data listed in Tables 1a and 1b. The complete selection comprises 1,000 documents per genre, thus enforcing equal prior classification probabilities for all genres.

We train genre classifiers with Support Vector Machines (SVM) with linear kernels, using the WEKA data mining software (Hall et al., 2009). As our features, we use the union of the 500 most common words per genre. We do not remove stopwords since they have a high potential to distinguish between various genres, which is long known in text genre classification literature [Karlgren and Cutting, 1994; Kessler et al., 1997; Stamatatos et al., 2000; Dewdney et al., 2001]. Using this classifier-feature combination, the classification accuracy on the documents in the test portion of our web-crawled corpora is 97.0%, 83.9%, and 88.1% for Arabic, Chinese, and English, respectively.

4.2. Genre adaptation experiments

Armed with accurate genre classifiers, we next classify for each document in the test set its genre, and guide it to the most appropriate SMT 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 document has unknown origin.

Figures 4a–4d show the translation quality in 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. We measure statistical significance with respect to the genre-agnostic baseline using approximate randomization (Riezler and Maxwell, 2005), reporting significant differences at the $p \leq 0.05$ ($\triangle$) level.

For Arabic→English and Chinese→English (Tables 4a and 4b 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 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 Chinese news documents classified as editorial improve with a different genre-optimized system, for example the Chinese news documents classified as editorial improve with an editorial system (§4.2).
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 corresponding genre-specific system. Note that the genre-agnostic baseline system is never recommended based on classifier predictions, despite sometimes being the best option.

Table 4c shows the end-to-end results for Bulgarian→English translation. The Bulgarian genre classifier achieves 100% accuracy on news and speech. The editorial test documents are all classified as news, which is advantageous for the SMT output quality. Genre predictions for the colloquial test documents are distributed 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 4a 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. 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.

5. Conclusions and future work

In this paper, we have presented parallel evaluation corpora covering four genres for four language pairs. We used these multi-genre benchmarks to show that BLEU differences between genres can be large and that, for all genres and all language pairs, translation quality improves when using four genre-optimized systems rather than a single genre-agnostic system. Finally, we trained and used classifiers to route test documents to the most appropriate genre systems, and showed that this setup can be used to successfully adapt SMT systems to four different genres, even for genres with no available parallel training data.

While experiments in this paper are limited to phrase-based SMT, they can also be applied to neural MT, for which curvature matters. While experiments in this paper are limited to phrase-based SMT, they can also be applied to neural MT, for which curvature matters.

Acknowledgements

This research was funded in part by NWO under project numbers 639.022.213, 612.001.218, and 639.021.646.

6. Bibliographical References


