Chapter 1

Introduction

Human languages are highly structured both from a syntactic and from a semantic point of view. This fundamental property makes it possible to efficiently convey an unlimited number of semantic concepts through natural language sentences. Crucially, multilingual data created by translation involve an additional layer of structure, which pivots between the syntactic and semantic patterns that appear in the different manifestations of human language.

In this thesis, we present methods to automatically learn phrase-based Statistical Machine Translation (SMT) models that assume a latent bilingual structure as their central modelling variable. Acknowledging that each language is strongly characterised by its individual structural properties, we aim to learn a bilingual structure that augments and supersedes its monolingual counterparts to bridge the gap between them. This structure is a latent one, because the translation data that we use to discover it do not explicitly identify it. The parallel corpora we use, consist of source sentences in the language we wish to translate from, paired with an existing human translation in the target language, but without any information on why the particular translation was chosen.

The goal of uncovering the hidden structure of translation is not new. On the contrary, it has formed the spearhead of Machine Translation (MT) research, right from the first steps of this field and up to this day. Already in the early days of MT, researchers strove to manually identify the latent patterns of translation, and encode them as a set of rules that governs the translation process. However, it was gradually recognised that the level of complexity of cross-language communication rendered this effort extremely difficult. Statistical Machine Translation aims to overcome the limitations of rule-based systems, through automatically learning bilingual correspondences between the source and target sentences from parallel corpora. All SMT models also assume an explicit or implicit latent structure in translation data, and one of the central problems in SMT is learning this bilingual structure using a parallel corpus.

Crucially, SMT research, in its large majority, has been fairly modest in the
kinds of structure assumed in its models, not daring to explore the complexity and richness of bilingual data. SMT models mostly stay close to the lexical surface to model translation from strings to strings, greatly trivialising the syntactic aspects of language. These oversimplifying assumptions allowed to avoid the learning challenges posed by more complex models of translation.

In this work, we move further than this and contribute methods to model and learn the latent structure of translation. We choose to face the problems that plagued previous efforts in this direction and propose solutions. We find that, to a large extent, these problems can be attributed to the sparse nature of translation data. As the models become more complex, naive learning algorithms are increasingly exposed to the danger of fixating on the particularities and the inherent noise of training data, crucially missing the opportunity to identify the underlying patterns.

We contribute a learning framework that addresses these issues, based on a long-established Machine Learning method: Cross-Validation. We show how this can be fused with the well-understood Maximum Likelihood Estimation (MLE) approach, to formulate a Cross-Validated MLE (CV-MLE) learning objective that directly aims to discover latent patterns that generalise well. We further provide the Cross-Validated EM algorithm, an instance of the equally well-understood Expectation-Maximization algorithm, to optimise parameters of models employing latent variables according to the CV-MLE criterion.

We subsequently apply our learning framework to induce, for the first time using a clear learning objective, translation models which capture the hierarchical, recursive structure of translation. Our method learns how to exploit monolingual syntactic structure to discover linguistically motivated translation patterns. We empirically show that our learnt models compare favourably to the state-of-the-art across multiple language pairs. In this way, we showcase how learning the structural aspects of translation can aid in delivering tangible improvement in translation performance.

In the rest of this chapter, we briefly introduce the three concepts which underlie this thesis.

1. We highlight the latent character of bilingual correspondence and consider what this implies for methods aspiring to automatically learn it.

2. We discuss different approaches to modelling translation structure and introduce the translation paradigms that we will employ in later chapters.

3. We consider the task of learning models assuming latent translation structure variables and discuss some of the challenges that we address in the rest of the thesis.

We close the introduction to this work with an overview of each of the chapters that follow.
1.1 The Latent Nature of Translation

We regard as translation structure the bilingual patterns that describe the correspondences between pairs of sentences in two languages, with each considered as the translation of the other. This structure identifies how the components of each sentence map to those of its translation counterpart. As such sentence components we might for example consider words, contiguous or discontinuous phrases, linguistic constituents or semantic units. Translation structures describe how these components correspond to each other, explaining the transformations taking place during the translation process.

In an SMT model trained from parallel corpora, the model variables corresponding to the translation structure are latent. The training corpora consist of whole source sentences each paired with their target language translations, without further annotation regarding how their sub-strings relate to each other. Even though sometimes, as is the case in this thesis, these training sentence-pairs might also be word-aligned\(^1\), we still cannot directly identify in the data other bilingual patterns, such as the clustering of words into phrase-pairs or the hierarchical correspondences between bilingual spans. A model assuming latent translation structure considers the values of the latent structure variables as missing from the training data; the problem of modelling them involves learning from incomplete data.

The training data provide no explicit clues on the form or properties of the hidden translation structure. It is the task of the modeller to define these by setting up the parts of the translation model space relating to the latent structural variables. Modelling options include choosing a word or phrase-based approach, assuming a flat structure directly over the lexical surface or a multilevel hierarchical structure, establishing a link between syntactic and translation analyses etc.

We believe that good choices related to the assumptions on the form of the latent translation structure, are those that lead to learning translation models which generalise well and translate adequately. This entails that the appropriateness of a model assuming a certain flavour of hidden translation structure must be evaluated in relation to the data that it will first train upon and those that it will later process, as well as the algorithmic context within which it will be employed. The learning algorithms which are used to train it, as well as the translation (decoding) apparatus that will be used to select translations for source sentences given the trained model, can also have a significant impact on the actual translation performance of the model. Furthermore, some translation models perform better in practice for certain language pairs, even for certain translation directions between them.

\(^1\)Word-aligned sentence-pairs include the word to word correspondences between the source and target sentences. These correspondences can be automatically identified by trained word-alignment models.
Language Sparsity  Irrespective of the particular choices involved, translation structure modelling seeks to take advantage of the inherent structural properties of monolingual source and target data to better model the correspondences between them. One might argue that in the face of the increasing availability of parallel training data, aiming to understand these correspondences is not necessary. As the size of the training data grows, there is a higher chance of retrieving from them the translations for large segments of test source sentences, eliminating the need to analyse how smaller fragments combine. However, such a view disregards the sparse nature of language. While extracting the translations of multi-word fragments from the training data has been shown to significantly raise translation quality (Och and Ney, 2004; Koehn et al., 2003; Chiang, 2005a) and the empirical part of this thesis uses solely such models, there is a limit on the extent that this can be applied to avoid modelling how these fragments combine. Even if we had access to a parallel corpus consisting of all the sentences on the world wide web and their translations, we would find it hard to match longer segments of yet unseen source sentences. This hardly relates solely to rare uses of language, but also for seemingly ‘normal’ segments of sentences such as the first four words of this sentence\(^2\). Irrespective of the size of our training data, the sparsity of natural language makes modelling the latent structural aspects of translation necessary, in order to produce fluent translations that convey meaning accurately.

1.2 Modelling Translation Structure

The development of translation models in the literature has proceeded in a step-wise fashion. Right from the beginning of SMT, the seminal work on the IBM Statistical Machine Translation models (Brown et al., 1993) was presented as a succession of translation models of increasing complexity. Formulating translation models is challenging and involves weighing together the perceived expressiveness of the models on the one hand, with the complexity of the computations involved and the machine learning challenges on the other.

From a probabilistic point of view, any translation model assumes a certain amount of structure between sentence-pairs, by preferring translations with certain properties (e.g. monotone translations that largely keep the word-order intact) over alternative ones. In this thesis, we focus on models assuming transla-

\(^2\)Searching the web for the phrase ‘this hardly relates solely’ returns zero matches on Google, while ‘this hardly’ and ‘relates solely’ returns hundreds of thousands of matches. To be able to translate the original four-word phrase adequately, a system having access to a hypothetical parallel corpus with the size of the web must still know how to combine together the translations of its two-word sub-phrases. The same applies for other phrases from this paragraph such as ‘web and their translations’, ‘translations of multi-word fragments’, ‘while extracting the translations’, ‘view disregards the sparse’ and many others. A lot of relatively short phrases from this thesis, like in any other natural language text, have never been formulated before.
1.2. Modelling Translation Structure

The value of such multi-word units had been already recognised in the original IBM SMT models (Brown et al., 1993). While these early models are widely referred to as ‘word-based’, a subset of them\(^3\) considers how single words produce multiple words as their translation, and models their tendency to cluster together as a phrase in the translated sentence. However, it was Phrase-Based SMT (Och et al., 1999; Koehn et al., 2003) that introduced modelling the phrase to phrase (i.e. many-to-many words), mapping between the source and target sentences.

**A Phrase-Based Approach** In this thesis we also follow a phrase-based approach to translation, considering the correspondences between both contiguous and discontiguous multi-word segments of sentences. Under this view, the translation structure for a sentence-pair, consisting of a source sentence and its target language translation, involves the following aspects:

**Phrase Segmentation** The structure must describe how the sentence-pair is segmented in phrase-pairs, where each target phrase in a phrase-pair is the translation of its source counterpart. These phrase-pairs can be contiguous (e.g. in lowercased English-French ⟨i am / je suis⟩), or discontiguous (e.g. ⟨not / ne . . . pas⟩). Each phrase-pair is considered atomic, i.e. it cannot be further analysed in terms of combining together smaller phrase or word-pairs.

**Reordering** The target parts of phrase-pairs are frequently reordered in relation to the order of the source phrases they are paired with. The translation structure must specify how the source and target parts of phrase-pairs are positioned in the source and target sentences respectively.

**Abstract Hierarchical Structure** Some of the models explored in this thesis explain the correspondence between the phrase-pairs of a sentence-pair in terms of an abstract hierarchical structure. This makes use of abstract (unlexicalised) categories, possibly linguistically motivated, which are combined together to form a hierarchical, recursive structure spanning across the sentence-pair. This structure might also describe the reordering patterns between the phrase-pairs.

**Hierarchical Modelling** As the thesis progresses, we focus on models assuming an abstract hierarchical translation structure that progressively gets more involved. These models will be based on the probabilistic Synchronous Context-Free Grammar formalism and its Inversion-Transduction Grammar subset (Wu,

---

\(^{3}\)IBM Model 3 introduces a ‘word fertility’ variable tracking the number of words produced as the translation of a single word. Models 4 and 5 further model the tendency of these multiple words originating as translations of the same word to cluster together.
Chapter 1. Introduction

This formalism extends the familiar concept of probabilistic Context-Free Grammars from the monolingual to the bilingual domain, to model pairs of strings instead of single sentences. It allows to both model discontinuous phrase-pairs as well as a hierarchical bilingual structure making use of abstract categories that are recursively expanded to derive a sentence-pair.

The introduction of synchronous grammars for SMT (Wu, 1997; Chiang, 2005a) cleared the way to take advantage of the inherent hierarchical structure of language in Machine Translation. This created the conditions to bring together the hierarchical, phrase-based modelling of MT with the existing thread of research exploring linguistic syntax-based SMT (Yamada and Knight, 2001; Galley et al., 2004). The result is work which explores hierarchical translation models driven by linguistic syntax for monolingual data.

This thesis concerns itself with all of the above models of translation structure. A comprehensive presentation of these models can be found in Chapter 2, while Chapters 4 to 6 examine our empirical work on learning progressively more complex models assuming a latent translation structure.

1.3 Learning Phrase-Based Translation Structure

Expectation-Maximization After establishing a certain translation model space, the next step involves estimating its parameters from the parallel training data. The introduction of SMT methods in terms of the IBM SMT models was based on employing the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to estimate translation model parameters according to a Maximum Likelihood Estimation (MLE) learning objective. Translation models are typically too complex to compute an MLE estimate analytically. Instead, the EM algorithm iteratively climbs the training data likelihood function producing a series of estimates, each further raising the data likelihood until convergence to a local optimum. The same methodology has been applied to other word-based SMT models such as the HMM alignment model (Vogel et al., 1996). However, as the translation models became more complex and especially after the introduction of Phrase-Based SMT, the weaknesses of naive applications of EM became evident, and researchers turned to heuristic, ad hoc estimators.

Learning Fragment Models The transition to phrase-based models brought with it new learning challenges. As we discuss in detail in Chapter 3, phrase-based SMT models belong to the wider family of Fragment Models (FMs), which was first introduced in the context of Data Oriented Processing (Scha, 1990; Bod, 1992; Bod et al., 2003). FMs model complex data by considering how these are composed from data fragments of arbitrary sizes, up to regarding a complete data
point as a single fragment. This modelling approach of FMs is extremely powerful, allowing to learn models with an arbitrary level of abstraction from the training data instances, leaving it to the learning algorithm to select the abstraction level which better generalises.

However, arriving at an abstraction level which will perform well when applying the model on yet unseen data is, almost by definition, extremely difficult for any learning objective or training algorithm based on the notion of ‘fitting’ the training data. The model space of FMs includes estimates which fit the training data so well that they essentially memorise them, while at the same time failing to anticipate novel data instances. Learning algorithms fitting the training set will return such degenerate estimates, leading to poor generalisation performance. Maximum Likelihood estimation and the EM algorithm fall in this category and for this reason, when applied straightforwardly, are of little use to estimate phrase-based SMT models and Fragment Models on the whole.

These issues led to the current trend of training phrase-based SMT models heuristically. Interestingly, these learning challenges are hardly new in the field of Natural Language Processing. They have been encountered before in the literature on the estimation of a natural language parsing model: Data Oriented Parsing (Prescher et al., 2004; Zollmann and Sima’an, 2006). In this thesis, we show how phrase-based SMT modelling is related to this prior literature and Fragment Modelling in general, and how all these models can benefit from the application of more appropriate learning approaches.

**Learning to Generalise with CV-EM** In order to formulate a solution, we will consider how these learning challenges touch upon a foundational problem in Machine Learning: addressing overfitting and estimating the potential of models to generalise, frequently understood in terms of the well-known bias-variance trade-off. Increasing the complexity of a model typically increases its ability to fit the training data, but also entails the danger of adapting to their particularities too closely, missing the underlying patterns. There has been a host of solutions proposed to alleviate this problem and find a good balance between fit and generalisation capacity. These include data-driven methods such as validation and cross-validation, methods employing Bayesian priors to counter overfitting or information theoretic approaches such as the Bayesian (Schwarz, 1978) and Akaike (Akaike, 1974) Information Criteria.

In this work, we revisit the problem of employing the EM algorithm for phrase-based and hierarchical translation models. We show that a heuristic solution to the estimation problems is not necessary and aim to unlock the potential of EM as an estimator for modern SMT models, by directly addressing the learning challenges involved. To do this, we opt for the data-driven Cross-Validation method. We integrate Cross-Validation within the Expectation-Maximization algorithmic framework to arrive at Cross-Validated EM (CV-EM): an instance of the EM al-
algorithm which optimises model parameters according to a Cross-Validated Maximum Likelihood Estimation (CV-MLE) objective. The application of the CV-EM algorithm, which we present in Chapter 3, will have a crucial role in our empirical work. It will contribute in leading estimation away from the overfitting hypotheses over the value of the latent translation structure variables, and towards regions of the parameter space which appear to generalise well according to a cross-validation criterion.

Other chapters consider the development of learning methodologies centred around the CV-EM algorithm for a series of latent translation structure models of increasing complexity. Learning contiguous Phrase-Based SMT models in Chapter 4 addresses the challenges of disambiguating the segmentation of sentence-pairs in phrase-pairs. Chapter 5 builds upon this to proceed to learn phrase-based models assuming a relatively simple hierarchical translation structure. Finally, Chapter 6 introduces a methodology to induce models using a linguistically motivated abstract translation structure, taking advantage of cues related to the syntactic structure of language to explain the correspondences between the two sides of bilingual data.

1.4 Thesis Overview

We close this introductory chapter with an overview of the rest of the thesis. For each chapter, we describe the relevant research context and delineate our contributions, our key empirical findings and conclusions.

Chapter 2: The Crossroads Between Machine Translation and Machine Learning

In this chapter, we follow the crossing paths of Statistical Machine Translation and Machine Learning. We start by examining some of the modelling paradigms that have been influential on SMT research, such as the noisy-channel approach of Shannon (Shannon, 1948), and examine the contrast between generative and discriminative modelling of translation. We also consider a categorisation of translation modelling frameworks, according to their approach on abstracting away from the lexical surface, the nature of the assumed latent variables and the learning methodology applied.

We continue with a presentation of the SMT modelling frameworks that are relevant to this work, such as the IBM word-based SMT models (Brown et al., 1993), Phrase-Based SMT models (Och et al., 1999; Koehn et al., 2003; Marcu and Wong, 2002) and Hierarchical SMT (Wu, 1997; Chiang, 2005a). We pay particular attention to the modelling concepts behind the latent translation variables that each of these models assumes, such as word and phrase alignments, segmentation, reordering and translation hierarchical structure. For every modelling framework,
1.4. Thesis Overview

we highlight its impact in the SMT literature and describe the learning challenges it introduced and how these were treated. In this way, we trace the progression of Machine Learning use in SMT research, which starts from the employment of well-founded estimation methods such as the EM algorithm, only to gradually resort to heuristic ad hoc solutions as the learning challenges mounted.

This thesis aspires to reconnect the learning methodology for modern phrase-based and hierarchical SMT models with the principled learning approaches employed in early work on SMT, in order to overcome the limits of the heuristic training methods. In the second part of the chapter, we build the theoretical background that will allow us to gain insights in the problems involved and that will provide the foundations for the learning methodology we propose to address them. We present the Expectation-Maximization algorithm, together with its crucial algorithmic and estimation properties. We proceed to examine the Bias-Variance decomposition of the Generalisation Error produced by a model and discuss the application of the Cross-Validation method to estimate it. The EM algorithm and Cross-Validation will form the two theoretical pillars under the novel learning algorithm we introduce in the next chapter: Cross-Validated EM.

Chapter 3: Fragment Models Estimation with the CV-EM Algorithm

The assumptions behind many modelling paradigms for complex, structured data, such as Markovian modelling or Bayesian Networks, can be understood to model data by examining how these are derived by combining together fixed-size data fragments. The Data Oriented Processing (DOP) paradigm (Scha, 1990; Bod and Scha, 1996) introduced the concept of Fragment Modelling: the derivation of data points from data fragments of arbitrary sizes, up to considering full data points themselves as single fragments. This is a modelling approach which is highly interesting for phrase-based SMT models assuming a latent translation structure. These too belong to the family of Fragment Models, with the data fragments for these models being contiguous or non-contiguous phrase-pairs, and the rest of the latent structure describing how these combine together.

In Chapter 3 we begin by examining the fragment-based DOP paradigm and its well-known implementation for natural language parsing, Data Oriented Parsing (Bod et al., 2003). We then abstract away from particular applications of DOP, to examine the implications of training Fragment Models using estimators which maximise model fit, such as Maximum Likelihood Estimation, extending earlier findings (Prescher et al., 2004; Zollmann and Sima’an, 2006). We consider why such training methods fail to produce estimates that generalise and discuss some of the alternatives proposed in prior literature.

In the second part of the chapter, we contribute a novel learning algorithm for Fragment Models: Cross-Validated Expectation-Maximization (CV-EM). Firstly,
we examine the pitfalls related to the step of formulating a Fragment Model from
the training corpus. During this step, copies of large segments of the training
corpus are essentially integrated in the model space. This makes trivial and
useless the crucial learning step of disambiguating between our hypotheses over
the values of latent model variables by fitting the training data, as the hypotheses
that will be preferred do nothing more than memorise the training corpus.

To address this, we introduce Cross-Validated Maximum Likelihood Estima-
tion (CV-MLE), an estimation objective which cross-validates the hypotheses
over the missing part of incomplete data to safeguard against hypotheses which
do not generalise. We show how CV-MLE crucially retains many of the desirable
estimation properties of plain MLE. We then contribute a practical implementa-
tion of the CV-MLE optimisation in terms of the CV-EM algorithm. We show
that CV-EM is a true instance of the Expectation-Maximization algorithm and
discuss its algorithmic and estimation properties and guarantees. We close with a
comparison of CV-EM to prior research in model estimation and with an overview
of what CV-EM has to offer for Fragment Model estimation.

Chapter 4: Learning Phrase-Pair Segmentation

Chapter 4 is the first of a series of three chapters which make up the second part of
this thesis. They present our contributions on the learning of three distinct SMT
model families of increasing complexity, each considering different assumptions
on the form of the hidden latent translation structure.

In this chapter, we begin by contributing a method to learn the conditional
translation probabilities of Phrase-Based SMT (PBSMT) models employing con-
tiguous phrase-pairs, as a replacement for the heuristic estimators that are typi-
cally used. These probabilities are the central probabilistic component of PBSMT
models and estimating their values essentially boils down to disambiguating how
sentence-pairs segment into contiguous phrase-pairs.

Prior research had shown that a Maximum-Likelihood estimation objective as
optimised by the EM algorithm performs considerably worse than the heuristic
estimators (DeNero et al., 2006). We argue that this is not surprising, by showing
that PBSMT models are instances of Fragments Models, and for this reason
inherit the estimation problems that plague this model family. Even though
the heuristic estimators already provide reasonable translation performance, we
motivate the need for a better founded estimation methodology and describe how
CV-EM can be applied instead to estimate the PBSMT model parameters.

Our approach is based on using the CV-EM algorithm to disambiguate sentence-
pair segmentation by maximising the Cross-Validated conditional likelihood of
each sentence in a sentence-pair given its counterpart, across both translation di-
rections. Our algorithm explores a binary segmentation space where each phrase-
pair either combines monotonically or swaps in relation to the neighbouring ones.
Cross-validating this hypothesis space over segmentations using CV-MLE and
CV-EM contributes in overcoming the overfitting tendency of MLE and arrive at estimates which generalise well.

We evaluate our approach against a baseline employing a heuristic estimator for translation from French and German on one side, to English on the other. We find that our estimator performs at least on a par with the heuristic one, with some configurations even performing slightly better. These experiments showcase how the theoretically appealing properties of CV-MLE and CV-EM translate in competitive empirical results. This finding essentially invalidates the need for a heuristic estimator, as previously justified in the face of no access to alternatives which perform at least equally well.

Chapter 5: Learning Stochastic Synchronous Grammars

The previous chapter already introduced the use of a binary segmentation space. However, as our aim was to estimate the parameters of a PBSMT model, the models we examined did not venture further than the lexical surface. In this thesis, Chapter 5 marks the transition from such models assuming a flat latent translation structure, towards models which consider translation as a recursive process. The models we will examine are centred around a hierarchical latent translation structure variable. Their formulation is based on the binary subset of the stochastic Synchronous Context-Free Grammars (SCFGs), an extension of the Context-Free Grammars for parallel strings, where every production’s right-hand side employs up to two bilingual non-terminals. These grammars combine the availability of algorithms to process them with a reasonable polynomial complexity, together with a high coverage of translation phenomena (Wu, 1997; Huang et al., 2009), with both features underlining their potential as foundations for formulating translation models.

Modelling with a synchronous grammar aims to take advantage of the recursive nature of language, as described by monolingual grammars, to capture the bilingual translation patterns. Still, the expressiveness of these models introduces new modelling and learning challenges.

Firstly, while the SCFG formalism seems superficially highly similar to its monolingual predecessor, by linking together the recursive structures of the source and target sentences, the result is more than the ‘sum’ of the two. It also specifies the syntactic element correspondences and the reordering patterns between the syntactic structures of the two languages. In this chapter, we discuss this, argue that an SCFG grammar must be designed with these issues in mind and contribute a design which addresses this, the ‘switch’ SCFG.

Secondly, the latent variable of stochastic SCFG models encapsulates several aspects of translation which previous models considered separately. While PBSMT models separate phrase segmentation from reordering, the latent hierarchical structure assumed by SCFG-based models must not only capture both of these core parts of the translation processes, but also their interdependence. As
Chapter 1. Introduction

an example, such a latent variable must regulate the trade-off between a detailed multi-level hierarchical structure and the memorisation of longer phrase-pairs: a sentence-pair segmented in few, long phrase-pairs necessitates a relatively shallow hierarchical structure to explain how these combine together.

Chapter 5 considers if the CV-EM algorithm is able to effectively learn such latent variables despite these issues. First, we describe how models employing a phrase-based SCFG as their backbone fall into the Fragment Models family, motivating the use of the CV-EM algorithm for their estimation, and then describe an implementation of CV-EM for SCFGs. We then consider learning stochastic grammars based on two SCFG designs, a simple one reminiscent of the abstract structure employed by standard hierarchical SMT implementations (Chiang, 2005a), as well as one employing our ‘switch’ SCFG design. We test the CV-EM induced grammars for both designs against a standard hierarchical translation baseline on a translation task from French to English. We find that both designs offer translation performance on a par with the heuristically estimated baseline, with the ‘switch’ SCFG scoring better than the simpler variation.

Chapter 5 is crucial in examining the potential of the CV-EM algorithm to learn translation models that assume a latent translation structure which, in contrast to Chapter 4, is not directly attached to the observed lexical surface. By confirming that our learning methodology is able to also learn such latent variables, we prepare the grounds for the next chapter. There, we transition from the simple hierarchical structures examined in Chapter 5 towards a significantly more complex, linguistically motivated latent translation structure.

Chapter 6: Learning Linguistically Motivated Latent Translation Structure

While structure can be found in natural language when examining it at different syntactic and semantic levels, linguistic syntax is widely considered as one of its most salient properties. For this reason, the linguistic structure of sentences has been targeted for more than a decade as an informative data source that can lead to better translations. This has further turned the spotlight towards methods which take a syntactic but not necessarily linguistic approach to translation, such as SCFG-based approaches, as promising devices to model the dependence of the translation process on linguistic notions of natural language structure.

Nevertheless, even though a host of translation phenomena can be described in linguistic terms, it must be recognised that, overall, linguistic structure correlates with a mere subset of the transformations that take place between the two sides of a language pair (Dorr, 1994; Fox, 2002; Koehn et al., 2003). As a result, methods which assume that translation can be fully explained in terms of correspondences and transformations that are solely driven by the linguistic structure of sentences, frequently fail to deliver competitive translation performance, due to imposing
unnecessary constraints on the translation process. The challenge is to find ways

In Chapter 6, we contribute a method to learn a linguistically motivated hierarchical translation model, by identifying the linguistic patterns which are informative for translation. We begin by constructing, for each training data sentence-pair, a chart covering with multiple linguistically motivated labels each aligned bilingual span. These labels are extracted from linguistic parses of the source sentence, where each of the multiple labels covering every span describes it from different linguistic perspectives and at varying levels of granularity. We then consider all binary structures which employ these labels to analyse the parallel training data, and use a translation-centric learning objective to disambiguate between them, according to their ability to explain the translation correspondences. This allows us to learn a model which is able to recursively analyse in linguistic terms the translation process across the whole sentence.

Our methodology builds on the foundations laid in the previous chapters. The synchronous recursive structure we consider, the Hierarchical-Reordering SCFG (HR-SCFG), is based on the principles behind the 'switch' SCFG of Chapter 5. The learning algorithm is an implementation of the Cross-Validation EM algorithm, as introduced in Chapter 3. In Chapters 4 and 5 we applied CV-EM for simpler translation models with most of their parameters directly relating to the lexical surface. Here, we separate the estimation of the lexical part of the model from the part related to the higher-level abstract hierarchical structure, and apply CV-EM to learn the latter: a linguistically motivated recursive structure which explains the correspondences and transformations between source and target sentences.

Crucially, contrary to other syntax-driven approaches (Way, 1999; Poutsma, 2000; Yamada and Knight, 2001; Galley et al., 2006; Huang et al., 2006; Liu et al., 2006), our method is linguistically motivated but not constrained. A translation-centric CV-MLE learning objective makes sure that only linguistically informed structures that help to explain translation are preferred, while the use of Cross-Validation aids in discovering those structures which are likely to generalise.

Other work (Marton and Resnik, 2008; Venugopal et al., 2009; Chiang et al., 2009) takes a more flexible approach, which is more similar to our own efforts. They opt to influence translation output using linguistically motivated features, or features based on source-side linguistically-guided latent syntactic categories (Huang et al., 2010). However, the features employed by these methods are local in nature, considering the linguistic plausibility of applying individual synchronous rules. As a result, these efforts totally lack the concept of a linguistically motivated hierarchical abstract structure reaching across the whole sentence-pair, which is exactly the focus of our own methodology.

The work of (Hassan et al., 2009) stands somewhat in the middle in comparison with fully syntax-driven SMT on the one hand and approaches using local
syntax-based features on the other. Their system extends the Direct Translation Model of Ittycheriah and Roukos (2007) with dependency-grammar based syntactic features, and takes under account an incrementally built target language dependency structure. However, their system solely considers minimal phrase-pairs which translate single source words, and, while they reach further than other feature-based systems, their target-side syntactic analyses are eagerly constructed without reference to a globally optimal structure. In contrast, we specifically focus on the challenges involved with training phrase-based systems with many-to-many phrase correspondences, and search for the bilingual structures that best explain sentence-pairs in their entirety.

We complete the picture, by contributing a set of decoding techniques to efficiently and effectively translate using the latent translation structure model learnt by CV-EM. We find that the learnt models and our translation system provides statistically significant translation improvements, up to +1.92 BLEU score points, for four different empirical tasks, translating from English to French, German, Dutch and Chinese.

The results of Chapter 6 complete those of Chapters 4 and 5, to provide considerable evidence to back the key hypothesis of this thesis: models assuming a latent translation structure can be learnt under a clear learning objective, as implemented in terms of a well-understood optimisation framework and learning algorithm. The learnt models are able to provide real-world, competitive translation performance in comparison to heuristic training regimes, rendering the use of the latter unnecessary. Still, we believe that the true potential of our methodology is not in providing a reliable and effective substitute for these heuristic estimators. On the contrary, it lies in carving a path to the future, by making possible the estimation of powerful translation models that uncover the latent side of translation, and whose estimation under ad hoc algorithms would have been hardly possible.
1.4. Thesis Overview

Sources of the Chapters

Some chapters of this dissertation are partially based on the following publications or present experimental results that were first reported in them. The Cross-Validated Expectation-Maximization algorithm presented in Chapter 3 has first appeared in (Mylonakis and Sima’an, 2008) and is further discussed in (Mylonakis and Sima’an, 2010). Chapter 4 is partially based on material and results first included in (Mylonakis and Sima’an, 2008). Chapter 5 is similarly related to (Mylonakis and Sima’an, 2010), while Chapter 6 extends material first published in (Mylonakis and Sima’an, 2011).

