Learning the latent structure of translation
Mylonakis, M.

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: http://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.
This dissertation discusses methods to learn the latent structural patterns that underlie translation data. It explores different approaches to modelling bilingual structure and presents novel frameworks and algorithms, such as Cross-Validated Expectation-Maximization (CV-EM), to learn phrase-based, hierarchical and syntax-driven Statistical Machine Translation (SMT) models from data.

In this thesis, we present methods to automatically learn phrase-based Statistical Machine Translation 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 by explaining the transformations taking place when conveying meaning across languages. The learning frameworks and algorithms we present allow us to discover these structural patterns in bilingual data and automatically learn models that take them into account to better translate. We apply our methodology for a sequence of statistical translation models of increasing complexity. This leads us to the presentation of a well-founded learning framework for hierarchical, syntactically motivated models that explain the translation process by taking advantage of the linguistic structure of language.

Chapter 1 offers an introduction to the context and aims of this work. It introduces the key aspects related to modelling translation structure and discusses the impact of its latent nature, as well as the challenges involved in learning to identify it in bilingual data. In Chapter 2, we start by examining some of the modelling frameworks that have been influential on SMT research, such as word-based, phrase-based and hierarchical SMT. We then discuss the EM algorithm and Cross-Validation, the two theoretical pillars under the novel learning algorithm we introduce in the chapter that follows. Chapter 3 examines the challenges related to learning phrase-based translation models, by considering the wider problem of learning Fragment Models: models which describe how to build new data instances by combining together data fragments extracted from a training
dataset. We then introduce the Cross-Validated Expectation-Maximization (CV-EM) algorithm, a novel learning algorithm for Fragment Models which optimises parameters according to a Cross-Validated Maximum Likelihood Estimation (CV-MLE) objective.

The next three chapters describe and empirically evaluate learning frameworks with CV-EM at their core, for three distinct, state-of-the-art SMT models. Chapter 4 contributes a well-founded method to learn the conditional translation probabilities of Phrase-Based SMT models employing contiguous phrase-pairs, centred around disambiguating the latent segmentation of sentence-pairs into phrase-pairs. This method is shown empirically to perform at least as well as the heuristic, ad hoc estimators that are typically used for these models. In Chapter 5, we consider the additional challenges involved in modelling translation with a synchronous grammar, and successfully learn a relatively simple hierarchical translation model which offers comparable performance with a highly competitive baseline. Chapter 6 moves considerably further, to build around CV-EM a learning framework that allows learning complex hierarchical translation models that take advantage of external annotations of source and/or target sentences. We deploy this framework to contribute a method to learn linguistically motivated hierarchical translation models, by identifying the source-language linguistic patterns which are informative for translation. We subsequently show how our approach delivers tangible translation improvements across four distinct language pairs.

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. Our methodology not only provides a reliable and effective substitute for these heuristic estimators, but most importantly lays 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.