Grammatical Inference for Syntax-Based Statistical Machine Translation

Menno van Zaanen[1] and Jeroen Geertzen[2]

[1] Division of Information and Communication Sciences
Department of Computing
Macquarie University
2109 Sydney, NSW, Australia
menno@ics.mq.edu.au
[2] Language and Information Science
Tilburg University
Tilburg, The Netherlands
j.geertzen@uvt.nl

Abstract. In this article we present a syntax-based translation system, called TABL (Translation using Alignment-Based Learning). It translates natural language sentences by mapping grammar rules (which are induced by the Alignment-Based Learning grammatical inference framework) of the source language to those of the target language. By parsing a sentence in the source language, the grammar rules in the derivation are translated using the mapping and subsequently, a derivation in the target language is generated. The initial results are encouraging, illustrating that this is a valid machine translation approach.

1 Introduction

Recently, there has been an increased interest in Statistical Machine Translation (SMT) [1]. SMT systems can be built using plain text only, which cuts down the development time of new MT systems immensely.

Some approaches that combine statistical learning with structured data by aligning syntax trees in two languages have been proposed previously [2, 3]. Here, we propose to use the Alignment-Based Learning framework [4] to generate these tree structures automatically.

For this task, the tree structures generated by ABL do not need to be linguistically correct. ABL only has to learn how words and phrases in one language (the source language) translate to words or phrases in another language (the target language).

2 Translation using Alignment-Based Learning

Translation using Alignment-Based Learning (TABL) automatically learns a machine translation system from a sentence aligned bi-lingual corpus. ABL is applied to sentences in the source language and their translation in the target

language. This structural information is used to analyse new source sentences and to generate translations.

TABL consists of two phases, the *training phase* and the *translation phase*. During the training phase ABL is applied to plain text translations in both the source and target language, creating a bracketed version of this data. Next, probabilistic context-free grammar (PCFG) rules are extracted [4, p. 53]. Grammar rules found in the derivations of both languages are stored, mapping each of the rules from the source language to the relevant (induced) rules in the target language. A concurrence score is also stored with each mapped rule.

During the translation phase, TABL parses a new source sentence using the PCFG of the source language. Next, using the mapping, the grammar rules of the derivation tree are mapped to those of the target language and a derivation in the target language is created. The yield of this derivation is the translation.

3 Results

To investigate how well this approach to machine translation really works, we applied TABL to different aligned corpora. Our test corpus showed over 80% correct translations, but about 50% of the sentences were not translated at all.

The reason why some sentences are not translated is that the target derivation is created by mapping from the source derivation. If the target derivation needs a different number of grammar rules, it cannot currently be generated.

4 Conclusion

In this article, we introduced TABL, a structure-based machine translation system, which demonstrates a novel application of the Alignment-Based Learning grammatical inference framework. Plain text collections that are aligned on sentence level are analysed, which results in grammars that show regularities in both languages. These regularities are then related between the languages. The resulting mapping illustrates how parts of sentences in the source language can be translated into equivalent parts in the target language.

Bibliography

[1] P. F. Brown, S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312, 1993.

[2] I. Dan Melamed. Statistical machine translation by parsing. In *42th Annual Meeting of the Association for Computational Linguistics; Barcelona, Spain*, 2004.

[3] Katharina Probst. *Automatically Induced Syntactic Transfer Rules for Machine Translation under a Very Limited Data Scenario*. PhD thesis, Carnegie Mellon University, Pittsburgh:PA, USA, 2005.

[4] Menno van Zaanen. *Bootstrapping Structure into Language: Alignment-Based Learning*. PhD thesis, University of Leeds, Leeds, UK, January 2002.