

## ABSTRACT

On this poster we present a syntax-based translation system, called TABL (Translation using Alignment-Based Learning). The system translates natural language sentences by first inducing grammar rules using Alignment-Based Learning and subsequently mapping the induced grammar rules from the source language to 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.

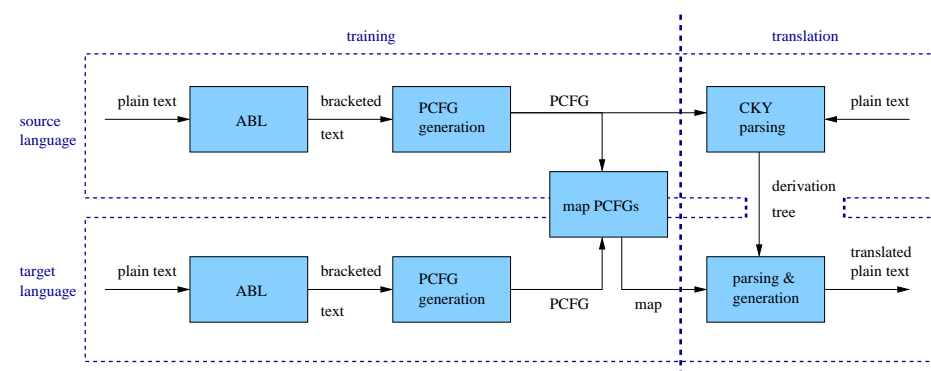
For this task, the tree structures generated by ABL do not need to be linguistically correct. ABL only has to learn how words and phrases in the source language translate to words or phrases in the target language.

## 2 Translation using Alignment-Based Learning

TABL automatically learns a machine translation system from a sentence aligned bilingual corpus. ABL is applied to sentences in the source language and their translation in the target language. The structural information induced is used to analyse new source sentences and to generate translations.

TABL consists of two phases: the **training** phase and the **translation phase**.

## 3 Overview TABL



## 4 Training 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:

Plain text		ABL output	
French	English	French	English
<i>je suis très malade</i>	<i>i am very sick</i>	<i>je suis (très malade)<sub>X</sub></i>	<i>i am (very sick)<sub>X</sub></i>
<i>je suis fatigué</i>	<i>i am tired</i>	<i>je suis (fatigué)<sub>Y</sub></i>	<i>i am (tired)<sub>Y</sub></i>

Next, probabilistic context-free grammar (PCFG) rules are extracted from the bracketed text. 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. A simplified example mapping could look as follows:

Mapping		
Source	Target	
$S_E \rightarrow NP_E VP_E$	$S_F \rightarrow NP_F VP_F$	1.00
$NP_E \rightarrow I_E$	$NP_F \rightarrow je_F$	1.00
$VP_E \rightarrow talk_E$	$VP_F \rightarrow parle_F$	0.75
$VP_E \rightarrow talk_E$	$VP_F \rightarrow parlons_F$	0.25
$VP_E \rightarrow talk_E$	$XP_F \rightarrow marche_F$	1.00

By considering cocurrence of the grammar rules in the derivations of the pairs of sentences, we can compute the likelihood of a grammar rule in the source language to translate into a particular grammar rule in the target language.

## 5 Translation phase

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.

For instance, the source sentence "I talk" is parsed according to the grammar. This results in the derivation rules as indicated in the table below in the source column.

Translating				
Source	Target 1		Target 2	
$S_E \rightarrow NP_E VP_E$	$S_F \rightarrow NP_F VP_F$	1.00	$S_F \rightarrow NP_F VP_F$	1.00
$NP_E \rightarrow I_E$	$NP_F \rightarrow je_F$	1.00	$NP_F \rightarrow je_F$	1.00
$VP_E \rightarrow talk_E$	$VP_F \rightarrow parle_F$	0.75	$VP_F \rightarrow parlons_F$	0.25
I talk	je parle	0.75	je parlons	0.25

For each of the grammar rules used in the derivation, corresponding translation rules are found. The first grammar rule only has one translation, but the third one has two. This results in two different possible translations (column Target 1 and Target 2). Note that the translation of " $VP_E \rightarrow talk_E$ " into " $XP_F \rightarrow marche_F$ " is not used, since "XP" does not fit in the partial derivation at that point. In the end, the two possible derivations in the target language are disambiguated by their probabilities. The translation "je parle" has a higher probability and is selected to be the translation.

## 6 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 reasons why these sentences are not translated are the following:

- The structures that ABL induces for the source and target language may not allow the generation of a target derivation, making it impossible for some cases to come up with a translation;
- The target derivation is created by mapping from the source derivation. If the target derivation requires a different number of grammar rules, it cannot be generated *at the moment*;
- The most probable derivation of the source sentence (generated by the PCFG parser) is not necessarily the derivation that would have been generated by ABL;

The current system can only find a translation for the cases where there is a derivation in the target language that has the same number of non-terminals as the most probable derivation in the source language. There may be several solutions to this problem, although we think that the easiest solution is to take a Data-Oriented Translation approach [5]. This allows not only CFG rules, but also larger parts of the derivation trees to map.

## 7 Conclusions

On this poster we described 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.

The initial results are promising in that many sentences (that have a translation) are actually translated correctly. However, about half of the sentences could not be translated. Future work will focus on allowing non-isomorphism in the mapping and prune away unlikely combinations, further reducing the effect of combinatorial explosion in the mapping combinations.

## References

- [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.
- [5] Arjen Poutsma. Data-Oriented Translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING); Saarbrücken, Germany*, pages 635–641. Association for Computational Linguistics, July 31–August 4 2000.