Syntax-based Statistical Machine Translation using Tree Automata and Tree Transducers

Daniel Emilio Beck
Computer Science Department
Federal University of São Carlos
daniel_beck@dc.ufscar.br

Abstract

In this paper I present a Master's thesis proposal in syntax-based Statistical Machine Translation. I propose to build discriminative SMT models using both *tree-to-string* and *tree-to-tree* approaches. Translation and language models will be represented mainly through the use of Tree Automata and Tree Transducers. These formalisms have important representational properties that makes them well-suited for syntax modeling. I also present an experiment plan to evaluate these models through the use of a parallel corpus written in English and Brazilian Portuguese.

1 Introduction

Statistical Machine Translation (SMT) has dominated Machine Translation (MT) research in the last two decades. One of its variants, Phrase-based SMT (PB-SMT), is currently considered the state of the art in the area. However, since the advent of PB-SMT by Koehn et al. (2003) and Och and Ney (2004), purely statistical MT systems have not achieved considerable improvements. So, new research directions point toward the use of linguistic resources integrated into SMT systems.

According to Lopez (2008), there are four steps when building an SMT system: *translational equivalence modeling*[1]*, parameterization, parameter estimation* and *decoding*. This Master's thesis proposal aims to improve SMT systems by including syntactic information in the first and second steps. Therefore, I plan to investigate two approaches: the Tree-to-String (TTS) and the Tree-to-Tree (TTT) models. In the former, syntactic information is provided only for the source language while in the latter, it is provided for both source and target languages.

There are many formal theories to represent syntax in a language, like Context-free Grammars (CFGs), Tree Substitution Grammars (TSGs), Tree Adjoining Grammars (TAGs) and all its synchronous counterparts. In this work, I represent each sentence as a constituent tree and use Tree Automata (TAs) and Tree Transducers (TTs) in the language and translation models.

Although this work is mainly language independent, proof-of-concept experiments will be executed on the English and Brazilian Portuguese (en-ptBR) language pair. Previous research on factored translation for this pair (using morphological information) showed that it improved the results in terms of BLEU (Papineni et al., 2001) and NIST (Doddington, 2002) scores, as shown in Table 1 (Caseli and Nunes, 2009). However, even factored translation models have limitations: many languages (and Brazilian Portuguese is not an exception) have relatively loose word order constraints and present long-distance agreements that cannot be efficiently represented by those models. Such phenomena motivate the use of more powerful models that take syntactic information into account.

2 Related work

Syntax-based approaches for SMT have been proposed in many ways. Some apply the TTS model: Yamada and Knight (2001) uses explicit inser-

[1] For the remainder of this proposal, I will refer to this step as simply *translation model*.

Proceedings of the ACL-HLT 2011 Student Session, pages 36–40,
Portland, OR, USA 19-24 June 2011.

	en-ptBR		ptBR-en	
	BLEU	NIST	BLEU	NIST
PB-SMT	0,3589	7,8312	0,3903	8,3008
FT	0,3713	7,9813	0,3932	8,4421

Table 1: BLEU and NIST scores for PB-SMT and factored translation experiments for the en-ptBR language pair

tion, reordering and translation rules, Nguyen et al. (2008) uses synchronous CFGs rules and Liu et al. (2006) uses TTs. Galley et al. (2006) also uses transducer rules but extract them from parse trees in target language instead (the *string-to-tree* approach - STT). Works that apply the TTT model include Gildea (2003) and Zhang et al. (2008). All those works also include methods and algorithms for efficient rule extraction since it's unfeasible to extract all possible rules from a parsed corpus due to exponential cost.

There have been research efforts to combine syntax-based systems with phrase-based systems. These works mainly try to incorporate non-syntatic phrases into a syntax-based model: while Liu et al. (2006) integrates bilingual phrase tables as separate TTS templates, Zhang et al. (2008) uses an algorithm to convert leaves in a parse tree to phrases before rule extraction.

Language models that take into account syntactic aspects have also been an active research subject. While works like Post and Gildea (2009) and Vandeghinste (2009) focus solely on language modeling itself, Graham and van Genabith (2010) shows an experiment that incorporates a syntax-based model into an PB-SMT system.

3 Tree automata and tree transducers

Tree Automata are similar to Finite-state Automata (FSA), except they recognize trees instead of strings (or sequences of words). Formally, FSA can only represent Regular Languages and thus, cannot efficiently model several syntactic features, including long-distance agreement. TA recognize the so-called Regular Tree Languages (RTLs), which can represent Context-free Languages (CFLs) since a set of all syntactic trees of a CFL is an RTL (Comon et al., 2007). However, it is important to note that the reciprocal is not true: there are RTLs that cannot be modeled by a CFL because those cannot capture the inner structure of trees. Figure 1 shows such an RTL, composed of two trees. If we extract an CFG from this RTL it would have the recursive rule S $\rightarrow$ SS, which would generate an infinite set of syntactic trees. In other words, there isn't an CFG capable to generate only the syntactic trees contained in the RTL shown in Figure 1. This feature implies that RTLs have more representational power than CFLs.

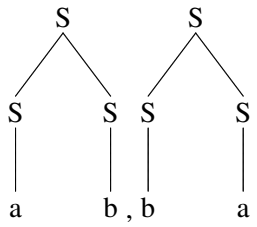

Figure 1: An RTL that cannot be modeled by a CFL

As a Finite-state Transducer (FST) is an extension of an FSA that produces strings, a Tree Transducer is an extension of a TA that produces trees. An FST is composed by an input RTL, an output RTL and a set of transformation rules. Restrictions can be added to the rules, leading to many TT variations, each with its properties (Graehl et al., 2008). The variations studied in this work are the *xT* (*extended top-down*, for TTT models) and *xTS* (*extended top-down tree-to-string*, for TTS models).

Top-down (T) transducers processes input trees starting from its root and descending through its nodes until it reaches the leaves, in contrast to *bottom-up* transducers, which do the opposite. Figure 2 shows a T rule, where uppercase letters (*NP*) represent symbols, lowercase letters (*q, r, s*) represent states and *x1* and *x2* are variables (formal definitions can be found in Comon et al. (2007)). Default top-down transducers must have only one symbol on the left-hand sides and thus cannot model some syntactic transformations (like local reordering, for example) without relying on copy and delete operations (Maletti et al., 2009). Extended top-down transducers allow multiple symbols on left-hand sides, making them more suited for syntax modeling. This property is shown on Figure 3 (adapted from Maletti et al. (2009)). Tree-to-string transducers simply drop the tree structure on right-

hand sides, which makes them adequate for translation models wihtout syntactic information in one of the languages. Figure 4 shows an example of a xTS rule, applied for the en-ptBR pair.

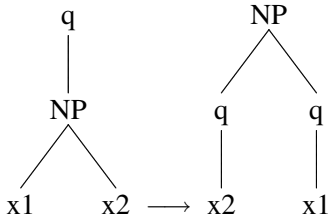

Figure 2: Example of a T rule

4 SMT Model

The systems will be implemented using a discriminative, log-linear model (Och and Ney, 2002), using the language and translation models as feature functions. Settings that uses more features besides those two models will also be built. In particular, I will investigate settings that incorporate non-syntactic phrases, using methods similar to Liu et al. (2006) and Zhang et al. (2008)

The translation models will be *weighted* TTs (Graehl et al., 2008), which add probabilities to the rules. These probabilities will be learned by an EM algorithm similar to the one described in Graehl et al. (2008). Rule extraction for TTS will be similar to the GHKM algorithm described in Galley et al. (2004) but I also plan to investigate the approaches used by Liu et al. (2006) and Nguyen et al. (2008). For TTT rule extraction, I will use a method similar to the one described in Zhang et al. (2008).

I also plan to use language models which takes into account syntactic properties. Although most works in syntactic language models uses tree grammars like TSGs and TAGs, these can be simulated by TAs and TTs (Shieber, 2004; Maletti, 2010). This property can help the systems implementation because it's possible to unite language and translation modeling in one TT toolkit.

5 Methods

In this section, I present the experiments proposed in my thesis and the materials required, along with the metrics used for evaluation. This work is planned to be done over a year.

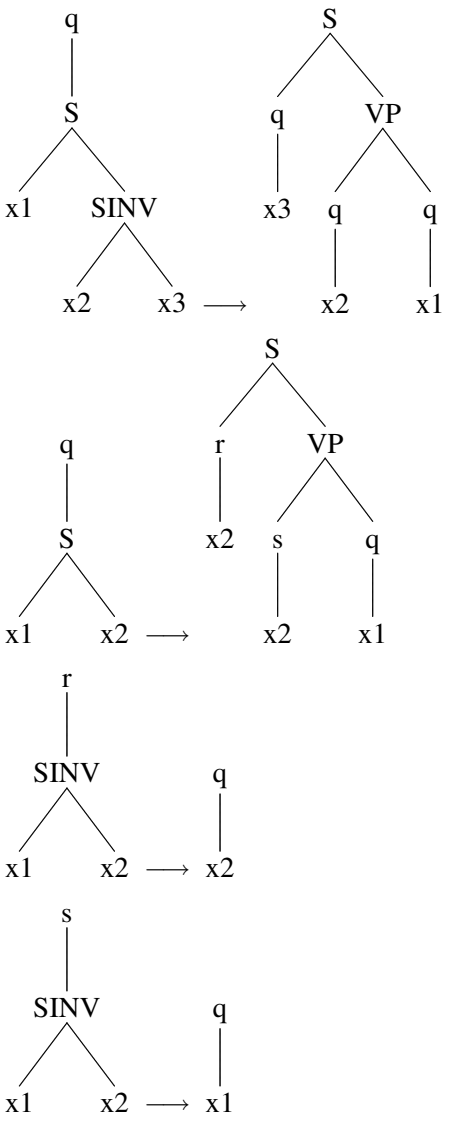

Figure 3: Example of a xT rule and its corresponding T rules

5.1 Materials

To implement and evaluate the techniques described, a parallel corpus with syntactic annotation is required. As the focus of this thesis is the English and Brazilian Portuguese language pair, I will use the PesquisaFAPESP corpus[2] in my experiments. This corpus is composed of 646 scientific papers, originally written in Brazilian Portuguese and manually translated into English, resulting in about 17,000 parallel sentences. As for syntactic annotation, I will use the Berkeley parser (Petrov and Klein, 2007) for

[2] http://revistapesquisa.fapesp.br

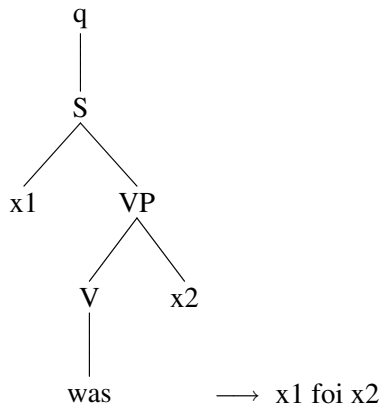

Figure 4: Example of a xTS rule (for the en-ptBR language pair)

English and the PALAVRAS parser (Bick, 2000) for Brazilian Portuguese.

In addition to the corpora and parsers, the following tools will be used:

- GIZA++[3] (Och and Ney, 2000) for lexical alignment
- Tiburon[4] (May and Knight, 2006) for transducer training in both TTS and TTT systems
- Moses[5] (Koehn et al., 2007) for decoding

5.2 Experiments and evaluation

Initially the corpus will be parsed using the tools described in section 5.1 and divided into a training set and a test set. For the TTS systems (one for each translation direction), the training set will be lexically aligned using GIZA++ and for the TTT system, its syntactic trees will be aligned using techniques similar to the ones proposed by Gildea (2003) and by Zhang et al. (2008). Both TTS and TTT systems will be implemented using Tiburon and Moses. For evaluation, BLEU and NIST scores on the test set will be used. The baseline will be the score for factored translation, shown in Table 1.

6 Contributions

After its conclusion, this thesis will have brought the following contributions:

- Language-independent SMT models which incorporates syntactic information in both language and translation models.
- Implementations of these models, using the tools described in Section 5.
- Experimental results for the en-ptBR language pair.

Technical reports will be written during this thesis progress and made publicly available. Paper submission showing intermediate and final results is also planned.

Acknowledgments

This research is supported by FAPESP (Project 2010/03807-4).

References

Eckhard Bick. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework.* Ph.D. thesis, Aarhus University.

Helena De Medeiros Caseli and Israel Aono Nunes. 2009. Tradução Automática Estatística baseada em Frases e Fatorada : Experimentos com os idiomas Português do Brasil e Inglês usando o toolkit Moses.

Hubert Comon, Max Dauchet, Remi Gilleron, Florent Jacquemard, Denis Lugiez, Christof Löding, Sophie Tison, and Marc Tommasi. 2007. *Tree automata techniques and applications*, volume 10. Available on: http://www.grappa.univ-lille3.fr/tata.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 128–132.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. Whats in a translation rule? In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*, pages 273–280.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 961–968.

[3] http://www.fjoch.com/GIZA++.html
[4] http://www.isi.edu/licensed-sw/tiburon
[5] http://www.statmt.org/moses

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 80–87.

Jonathan Graehl, Kevin Knight, and Jonathan May. 2008. Training Tree Transducers. *Computational Linguistics*, 34:391–427.

Yvette Graham and Josef van Genabith. 2010. Deep Syntax Language Models and Statistical Machine Translation. In *SSST-4 - 4th Workshop on Syntax and Structure in Statistical Translation at COLING 2010*, page 118.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 609–616.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):1–49.

Andreas Maletti, Jonathan Graehl, Mark Hopkins, and Kevin Knight. 2009. The power of extended top-down tree transducers. *SIAM Journal on Computing*, 39(2):410–430.

Andreas Maletti. 2010. A Tree Transducer Model for Synchronous Tree-Adjoining Grammars. *Computational Linguistics*, pages 1067–1076.

Jonathan May and Kevin Knight. 2006. Tiburon : A Weighted Tree Automata Toolkit. *Grammars*.

Thai Phuong Nguyen, Akira Shimazu, Tu-Bao Ho, Minh Le Nguyen, and Vinh Van Nguyen. 2008. A tree-to-string phrase-based model for statistical machine translation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning - CoNLL '08*, pages 143–150.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 295.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*, pages 404–411.

Matt Post and Daniel Gildea. 2009. Language modeling with tree substitution grammars. *Computing*, pages 1–8.

Stuart M Shieber. 2004. Synchronous Grammars as Tree Transducers. *Applied Sciences*, pages 88–95.

Vincent Vandeghinste. 2009. Tree-based target language modeling. In *Proceedings of EAMT*, pages 152–159.

Kenji Yamada and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. In *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530.

Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proc. ACL-08: HLT*, pages 559–567.