Using Parsed Corpora for Estimating Stochastic Inversion Transduction Grammars

Germán Sanchis-Trilles, Joan Andreu Sánchez

Instituto Tecnológico de Informática
Universidad Politécnica de Informática
Camino de Vera, s/n. 46022 Valencia, Spain
{gsanchis,jandreu}@dsic.upv.es

Abstract

An important problem when using Stochastic Inversion Transduction Grammars is their computational cost. More specifically, when dealing with corpora such as Europarl only one iteration of the estimation algorithm becomes prohibitive. In this work, we apply a reduction of the cost by taking profit of the bracketing information in parsed corpora and show machine translation results obtained with a bracketed Europarl corpus, yielding interresting improvements when increasing the number of non-terminal symbols.

1. Introduction

Statistical Machine Translation (SMT) systems have proved in the last years to be an important alternative to rule-based MT systems, being even able of outperforming commercial machine translation systems in the tasks they have been trained on. Phrase-based (PB) models (Tomas and Casacuberta, 2001; Zens et al., 2002) have proved to provide a very efficient framework for SMT.

An important issue when training PB models is the algorithm by means of which the bilingual phrases are extracted. Hence, a wide variety of methods have been proposed for this purpose, spanning through statistically motivated procedures (Tomas and Casacuberta, 2001), heuristic algorithms (Zens et al., 2002), and linguistically motivated methods (Sánchez and Benedí, 2006a). In this work, we will be following this last approach, which relies on Stochastic Inverse Transduction Grammars (SITGs) (Wu, 1997) for phrase extraction.

SITGs constitute a restricted subset of syntax directed stochastic grammars for translation, and are very related to context-free grammars. These can be used to analyse two strings simultaneously, which makes them specially useful for extracting bilingual segments from a parallel corpus in a syntax-oriented manner. In (Sánchez and Benedí, 2006b), SITGs were used for obtaining word phrases, reporting preliminary results on the EuroParl corpus. In this work, we extend that work by using bracketed corpora for estimating the STIGs.

In section 2, we will briefly review the phrase-based SMT approach. Next, in section 3, we will sum up the grounds of SITGs and the modifications proposed in (Sánchez and Benedí, 2006a). In section 4, we present the translation results on the Europarl corpus, obtained when applying one learning iteration on SITGs with several number of non-terminals.

2. Phrase-Based models

The derivation of the PB models stems from the concept of bilingual segmentation, i.e. sequences of source words and sequences of target words. It is assumed that only segments of contiguous words are considered, the number of source segments being equal to the number of target segments and each source segment being aligned with only one target segment and vice versa.

Ultimately, when learning a PB model, the purpose is to compute a *phrase translation table*, where each input phrase is assigned to one or more output phrases with a given probability. In this work, we use SITGs to build this phrase translation table.

3. Stochastic Inversion Transduction Grammars

Being closely related to context-free grammars, Stochastic Inverse Transduction Grammars (Wu, 1997) specify a subset of syntax directed stochastic grammars for translation. Analysing two strings simultaneously, SITGs may be used to extract bilingual segments from a parallel corpus while taking into account syntax-motivated restrictions. The internal nodes of the parse tree define a span over each pair of strings. These spans can be considered as paired segments of words.

In (Wu, 1997), an algorithm similar to the CYK of context free grammars is proposed in order to parse a sentence pair with a SITG. This algorithm has a time complexity of $O(|x|^3|y|^3|R|)$, being $|x|$ the length of the source sentence, $|y|$ the length of the target target sentence, and $|R|$ the number of rules in the SITG. However, if the corpus has been previously parsed with a syntactical parser and is given in a bracketed form, (Sánchez and Benedí, 2006a) suggest the use of a version of the algorithm by (Wu, 1997) which is more efficient while performing the analysis, achieving a time complexity of $O(|x||y||R|)$ when x and y are fully bracketed. In this work, we will be taking profit of bracketing information provided by freely available parsing toolkits in order to achieve an important increase of speed within the estimation algorithm.

4. Experiments

We performed our experiments on the Spanish-English Europarl corpus, with the partition established in the Workshop on Statistical Machine Translation of the NAACL 2006 (Koehn and Monz, 2006).

Table 1: Translation results for a SITG with only one, two, three and four non-terminal symbols. Results are shown in BLEU/WER. 0 iterations means the SITG was obtained by the heuristic technique.

non terms	It. 0	It. 1
1	26.8/62.5	26.9/62.6
2	27.0/62.6	27.5/62.1
3	26.9/62.7	27.0/62.7
4	26.6/63.2	27.9/61.5

4.1. SITGs for phrase extraction

First, we built an initial SITG by following the method described in (Sánchez and Benedí, 2006b). Then, both source and target languages in the training corpus were bracketed by using FreeLing (Asterias et al., 2006), which is an open-source suite of language analysers. This being done, we then used the bracketed corpus to perform one estimation iteration on the initial SITG and obtain improved SITGs. Finally, the SITG obtained after the estimation iteration was used to parse the bracketed training corpus and extract segment pairs to setup a phrase-based translation model.

It is important to stress the importance of the bracketing information, without which it would have been practically impossible to perform any learning iterations at all because of the severe temporal issues.

Following common knowledge in SMT, we computed both the inverse and direct translation probabilities of each segment pair according to the formulae

$$p(\mathbf{s}|\mathbf{t}) = \frac{C(\mathbf{s},\mathbf{t})}{C(\mathbf{t})} \qquad p(\mathbf{t}|\mathbf{s}) = \frac{C(\mathbf{s},\mathbf{t})}{C(\mathbf{s})}$$

where $C(\mathbf{s},\mathbf{t})$ is the number of times segments **s** and **t** were extracted throughout the whole corpus. This phrase-table was fed to Moses (Philipp Koehn, 2007) for producing the final translation.

Initial SITGs with increasing number of non-terminal symbols were built and then estimated. The purpose of building SITGs with several non-terminal symbols was to analyse whether augmenting the number of non-terminals would improve word reorderings between both input and output languages. Adding non-terminal symbols may provide more complexity to the grammar built, and hence increases its expressive power. (Sánchez and Benedí, 2006b)

Translation results of this setup can be seen in Table 1. Here, all the weights of the log-linear model were adjusted my MERT training, and the language model used was a 5-gram interpolated with Knesser-Ney discount.

It is interresting to point out that one estimation iteration for any number of non-terminal symbols has deffinitely an improving effect.

(Sánchez and Benedí, 2006b) shows an experiment in which segments were extracted from training corpora without any bracketing information. Since this was not computationally feasible with the training algorithm, we decided to use the SITG obtained after one estimation iteration (estimated using the bracketed corpus) to parse a non-bracketed version of the corpus for the purpose of obtaining segments. Interestingly however, the BLEU score did not vary. The same conclusion was achieved when mixing the segments obtained when using the bracketed and the non-bracketed corpus: again, the BLEU score did not differ significantly. The fact that the translation quality is not lessened by introducing the bracketing, and hence constraining the SITG, has a lot of importance, since a bracketed corpus can be analysed by the SITG much faster.

Table 2: Translation results for a SITG with only one, two, three and four non-terminal symbols when adding the syntactic models described. Results are shown in BLEU/WER.

non terms	It. 1	+ syntactic
1	26.9/62.6	27.7/61.6
2	27.5/62.1	28.3/61.1
3	27.0/62.7	28.2/61.3
4	27.9/61.5	28.9/60.0

4.2. Adding Syntactic Translation Probabilities

In the process of obtaining the best parse tree $\widehat{t}_{x,y}$ for each pair of strings (x, y) (see Section 3), a joint probability $\widehat{p}(\mathbf{s},\mathbf{t})$ (**s** and **t** are, respectively, word segments from x and y) for several overlapping spans is obtained. It is important to note that a given pair of word segments $(\mathbf{s},\mathbf{t})$ can have different probabilities depending on the tree it comes from. We have defined a new translation model that is based in this information as follows. Let Ω the multiset of spans (word segments) obtained from the training sample, and $\Omega_{\mathbf{s},\mathbf{t}} \subseteq \Omega$ the multiset of $(\mathbf{s},\mathbf{t})$ spans. We define the expected value of $\widehat{p}(\mathbf{s},\mathbf{t})$ according to the empirical distribution as:

$$E_\Omega(\widehat{p}(\mathbf{s},\mathbf{t})) = \frac{\sum_{(\mathbf{a},\mathbf{b})\in\Omega_{\mathbf{s},\mathbf{t}}} \widehat{p}(\mathbf{a},\mathbf{b})}{|\Omega|}.$$

If we marginalise for the input side of the word segments and for the output side of the segments, then we get:

$$E_\Omega(\widehat{p}(\mathbf{s})) = \sum_{\mathbf{t}} E_\Omega(\widehat{p}(\mathbf{s},\mathbf{t}))$$

and

$$E_\Omega(\widehat{p}(\mathbf{t})) = \sum_{s} E_\Omega(\widehat{p}(\mathbf{s},\mathbf{t})).$$

In this way we obtain these two new *syntax-based* models:

$$p(\mathbf{s}|\mathbf{t}) = \frac{E_\Omega(\widehat{p}(\mathbf{s},\mathbf{t}))}{E_\Omega(\widehat{p}(\mathbf{t}))}, \quad p(\mathbf{t}|\mathbf{s}) = \frac{E_\Omega(\widehat{p}(\mathbf{s},\mathbf{t}))}{E_\Omega(\widehat{p}(\mathbf{s}))}.$$

As can be seen in Table 2, adding these new syntax based models produces a consistent improvement of approximately one point of BLEU.

5. Discussion

Comparatively, the best result that (Sánchez and Benedí, 2006b) reported in the Spanish-English task was a BLEU score of 23.0, which they obtained by combining segments extracted from both the bracketed and the non-bracketed corpus. We have widely exceeded this baseline.

On the other hand, the Moses toolkit (Philipp Koehn, 2007), which is a state of the art statistical machine translation system, obtains in this task a score of 31.0 BLEU. However, when constrained to use only the inverse and direct translation models as we did, the score drops to 29.6 BLEU, which is only 1.7 points away from our best score, with only the direct and the inverse translation models, and 0.7 points away from our overall best score. It can be argued that we should be comparing this last score with Moses with all four models, adding the lexical alignment models. However, these lexical alignment models can also be added to our system, and we actually plan to do so as future work, whereas the syntactic models we introduced cannot be added to the segments obtained in the Moses Toolkit.

Although Moses obtains a slightly better score, it must be taken into consideration that this toolkit achieves this by using 19M different segment pairs, whereas our translation models only use half that amount. This fact has important implications: being our model smaller, less computational resources are used in decoding time, but also the final translation is produced faster.

Moreover, adding non-terminal symbols seems to have beneficial effects on the final BLEU score. Hence, it seems there is still room for improvement, whereas regular phrase-based models (such as Moses) do not have this ability.

6. Conclusions and Future Work

We have presented an alternative method for phrase extraction, which is competitive in terms of quality and produces smaller phrase-based models when compared to the traditional phrase-based extraction algorithms used.

Moreover, we have shown that freely available natural language processing toolkits can be successfully used to obtain bracketed corpora and reduce time complexity in SITG estimation, without trading off translation quality.

In the future, we plan to compute more complex SITGs and introduce further models to improve our translation table, such as the lexical alignment models or other models obtained by combining the various probabilities that SITG estimation entails. In this line, we also plan to investigate which effect has the combination of our phrase table with the phrase table produced by Moses.

7. Acknowledgements

This work has been partially supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, by the MEC scholarship AP2005-4023 and by the Universidad Politécnica de Valencia with the ILELA project.

8. References

J. Asterias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Alexandra Birch Chris Callison-Burch Marcello Federico Nicola Bertoldi Brooke Cowan Wade Shen Christine Moran Richard Zens Chris Dyer Ondrej Bojar Alexandra Constantin Evan Herbst Philipp Koehn, Hieu Hoang. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, demonstration session*.

J.A. Sánchez and J.M. Benedí. 2006a. Obtaining word phrases with stochastic inversion transduction grammars for phrase-based statistical machine translation. In *Proc. 11th Annual conference of the European Association for Machine Translation*, pages 179–186, Oslo, Norway, June.

J.A. Sánchez and J.M. Benedí. 2006b. Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 130–133, New York City.

J. Tomas and F. Casacuberta. 2001. Monotone statistical translation using word groups. In *Proceedings of the Machine Translation Summit VIII*, pages 357–361, Santiago de Compostela, Spain.

D. Wu. 1997. Stochastic iversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI. Lecture Notes in Computer Science*, volume 2479, pages 18–32.