PHRASE SEGMENTS OBTAINED WITH STOCHASTIC INVERSION TRANSDUCTION GRAMMARS FOR SPANISH-BASQUE TRANSLATION

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

One of the weaknesses of the so-called phrase-based translation models is that they carry out a blind extraction of the phrase translation table, i.e., they do not take into account the possible linguistic restrictions that each language introduces because of its own syntax. In this work, we use Stochastic Inversion Transduction Grammars as a phrase extraction technique which is able to yield similar results to more popular, but heuristic, techniques. We present encouraging results obtained on the Albayzin 2008 corpus.

1. INTRODUCTION

The grounds of modern Statistical Machine Translation (SMT), a pattern recognition approach to Machine Translation, were established in [1], where the problem of machine translation was defined as following: given a sentence $\mathbf{x}$ from a certain source language, an adequate sentence $\hat{\mathbf{y}}$ that maximises the posterior probability is to be found. Such a statement can be specified with the following formula

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, Pr(\mathbf{y}|\mathbf{x}). \tag{1}$$

Applying the Bayes theorem on this definition and operating appropriately, one can easily obtain the following formula

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}}\, Pr(\mathbf{y}) \cdot Pr(\mathbf{x}|\mathbf{y}), \tag{2}$$

where $Pr(\mathbf{y}|\mathbf{x})$ has been decomposed into two different probabilities: the *statistical language model* of the target language $Pr(\mathbf{y})$ and *the (inverse) translation model* $Pr(\mathbf{x}|\mathbf{y})$, and the denominator has been neglected because it does not affect the maximisation.

In practise, the direct modelling of the posterior probability $Pr(\mathbf{y}|\mathbf{x})$ has been widely adopted. To this purpose, different authors [2, 3] propose the use of the so-called log-linear models, where the decision rule is given by the expression

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \tag{3}$$

where $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, M is the number of models (or features) and λ_m are the weights of the log-linear combination.

2. PHRASE-BASED MODELS

The derivation of the Phrase-Based (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.

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 [4], heuristic algorithms [5], and linguistically motivated methods [6]. In this work, we will be following this last approach, which relies on Stochastic Inverse Transduction Grammars (SITGs) [7] for phrase extraction.

In this work we will be following the approach by [8], in which SITGs are used for phrase extraction, reporting preliminary results on the EuroParl corpus. In [9], such work was extended with a more thorough experimentation, improving considerably the translation quality previously obtained.

This work has been partially supported by the Spanish MEC under scholarship AP2005-4023 and under grants CONSOLIDER Ingenio-2010 CSD2007-00018, by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, and by the Generalitat Valenciana under grant GVPRE/2008/331, research project "Traducción Automótica del Corpus UPenn Treebank mediante Tócnicas Interactivas (UPennSpanish)."

3. STOCHASTIC INVERSION TRANSDUCTION GRAMMARS

Being closely related to stochastic context free grammars, Stochastic Inverse Transduction Grammars [7] specify a subset of stochastic syntax-directed stochastic grammars. 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 [7], an algorithm similar to the CYK algorithm for 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 input part of the corpus (the source language), the output part (the target language) or both of them has been previously parsed (each part with a monolingual parser) and is given in a bracketed form, [6] suggests the use of a version of the algorithm given in [7] 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 monolingual parsing toolkits in order to achieve an important increase of speed within the estimation algorithm, without a significant loss in terms of final translation quality [9].

4. SITGS FOR PHRASE EXTRACTION

First, we built an initial SITG by following the method described in [8]. The basic idea is to construct the maximum number of syntactic rules with a given number of non-terminal symbols. These non-terminal symbols were not syntacticly motivated. The lexical rules of the initial SITG were obtained from a lexical dictionary. Then, the source language in the training corpus (Spanish) was bracketed by using FreeLing [10], which is an open-source suite of language analysers. This being done, we then used the bracketed corpus to perform two stochastic estimation iterations on the initial SITG and obtain improved SITGs. Finally, the SITG obtained after the estimation iterations was used to parse the bracketed training corpus and extract segment pairs to setup a phrase-based translation model.

Once extracted, the phrase pairs were scored according to the following translation models:

1. 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.

2. We also scored the phrase pairs with syntax-based translation models. These are obtained following the technique described in [9], where each segment pair is assigned a probability according to the corresponding SITG. When a given segment pair $(\mathbf{s},\mathbf{t})$ is parsed by the SITG, a joint probability $\widehat{p}(\mathbf{s},\mathbf{t})$ is obtained. Since this probability may differ depending on the parse tree it comes from, we need to normalise accordingly. 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. The expected value of $\widehat{p}(\mathbf{s},\mathbf{t})$ is defined according to the empirical distribution as:

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

Similarly,

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

3. In addition, we also considered the use of lexical weights, as described in [5]. These lexical weights attempt to account for the lexical soundness of each phrase pair, estimating how well each of the words in one language translates to each of the words in the other language.

With these scores, we build three sets of phrase-tables. The first one was built by only including the direct and inverse translation probabilities (1) and the syntactic probabilities (2), since this was the combination reported in [9]. This combination will be referred as *Vsyn*. However, in [9], lexical weights were not included. For this reason, in these experiments we analysed the effect of only including direct and inverse translation probabilities and lexical weights (this combination will be referred to as Vlex), and including all six sets of probabilities (from now on, VII). These phrase-tables were fed to Moses [11] for producing the final translation.

5. EXPERIMENTS

We performed our experiments on the Spanish-Basque Albayzin corpus, with the partition established in the *V Jornadas en Tecnología del Habla* (2008). The statistics of the corpus can be seen on Table 1. As it can be seen on the Table, translating both from or into Basque is a difficult task, since the amount of Out of Vocabulary words quickly becomes very high.

As Table 2 shows, the translation quality tends to get better when increasing number of non-terminal symbols are used, as measured by BLEU. Moreover, the VII combination, in which all translation models are used, seems

Table 1. Characteristics of Albayzin corpus. OoV stands for "Out of Vocabulary" words, Dev. for Development, K for thousands of elements and M for millions of elements.

		Spanish	Basque
Training	Sentences	58K	
	Run. words	1151K	885M
	Avg. length	19.8	15.2
	Voc.	49.4K	87.8K
Dev.	Sentences	1456	
	Run. words	29K	23K
	Avg. length	20.1	15.5
	OoV	489	8376
Test	Sentences	1446	
	Run. words	28K	22K
	Avg. length	19.3	14.9
	OoV	483	8096

Table 2. Translation results for Spanish-Basque translation when using a SITG with only one, three and five non-terminal symbols

non terms	combination	BLEU	WER	TER
1	Vsyn	8.8	82.0	78.5
	Vlex	8.8	81.8	78.2
	VII	9.0	81.7	78.1
3	Vsyn	8.9	81.9	78.6
	Vlex	8.9	81.8	78.3
	VII	9.1	**81.4**	**77.9**
5	Vsyn	9.1	82.2	78.7
	Vlex	9.2	81.5	78.9
	VII	**9.3**	81.6	78.1

to yield improvements over the other alternatives, as measured by BLEU, WER and TER. However, it must be noted that these differences are not statistically significant. The results shown in this table were obtained restricting the decoder to perform a monotonic translation procedure, since at this stage we have not yet implemented a SITG-based reordering model. In this case, the language model used was a 5-gram, applying interpolation with Knesser-Ney discount.

For comparison purposes, the best scores obtained by the Moses toolkit in its monotonic setup are 9.4 BLEU, 81.7 WER and 78.3 TER, which are not significantly better than the scores obtained by our system trained with 5 non-terminal symbols in the VII combination.

6. CONCLUSIONS AND FUTURE WORK

We have presented an alternative method for phrase extraction, which is competitive in terms of quality. This method obtains phrase segments from paired sentences by parsing both of them in a completely unlexicalixed manner.

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. BIBLIOGRAFA

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, y Robert L. Mercer, "The mathematics of machine translation," in *Computational Linguistics*, June 1993, vol. 19, pp. 263–311.

[2] K. Papineni, S. Roukos, y T. Ward, "Maximum likelihood and discriminative training of direct translation models," in *Proc. of ICASSP'98*, 1998, pp. 189–192.

[3] F. Och y H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the ACL'02*, 2002, pp. 295–302.

[4] J. Tomas y F. Casacuberta, "Monotone statistical translation using word groups," in *Proceedings of the Machine Translation Summit VIII*, Santiago de Compostela, Spain, 2001, pp. 357–361.

[5] R. Zens, F.J. Och, y H. Ney, "Phrase-based statistical machine translation," in *Advances in artificial intelligence. 25. Annual German Conference on AI. Lecture Notes in Computer Science*, 2002, vol. 2479, pp. 18–32.

[6] J.A. Sánchez y J.M. Benedí, "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*, Oslo, Norway, June 2006, pp. 179–186.

[7] D. Wu, "Stochastic iversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[8] J.A. Sánchez y J.M. Benedí, "Stochastic inversion transduction grammars for obtaining word phrases for phrase-based statistical machine translation," in *Proceedings of the Workshop on Statistical Machine Translation*, New York City, 2006, pp. 130–133.

[9] G. Sanchis-Trilles y J.A. Sánchez, "Using parsed corpora for estimating stochastic inversion transduction grammars," in *6th edition of the International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 26 – June 1 2008.

[10] J. Asterias, B. Casas, E. Comelles, M. González, L. Padró, y M. Padró, "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, 2006.

[11] 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, "Moses: Open source toolkit for statistical machine translation," in *ACL 2007, demonstration session*, 2007.