Stochastic Inversion Transduction Grammars for Obtaining Word Phrases for Phrase-based Statistical Machine Translation

J.A. Sánchez and J.M. Benedí
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Valencia, Spain
jandreu@dsic.upv.es jbenedi@dsic.upv.es

Abstract

An important problem that is related to phrase-based statistical translation models is the obtaining of word phrases from an aligned bilingual training corpus. In this work, we propose obtaining word phrases by means of a Stochastic Inversion Translation Grammar. Experiments on the shared task proposed in this workshop with the Europarl corpus have been carried out and good results have been obtained.

1 Introduction

Phrase-based statistical translation systems are currently providing excellent results in real machine translation tasks (Zens et al., 2002; Och and Ney, 2003; Koehn, 2004). In phrase-based statistical translation systems, the basic translation units are word phrases.

An important problem that is related to phrase-based statistical translation is to automatically obtain bilingual word phrases from parallel corpora. Several methods have been defined for dealing with this problem (Och and Ney, 2003). In this work, we study a method for obtaining word phrases that is based on Stochastic Inversion Transduction Grammars that was proposed in (Wu, 1997).

Stochastic Inversion Transduction Grammars (SITG) can be viewed as a restricted Stochastic Context-Free Syntax-Directed Transduction Scheme. SITGs can be used to carry out a simultaneous parsing of both the input string and the output string. In this work, we apply this idea to obtain aligned word phrases to be used in phrase-based translation systems (Sánchez and Benedí, 2006).

In Section 2, we review the phrase-based machine translation approach. SITGs are reviewed in Section 3. In Section 4, we present experiments on the shared task proposed in this workshop with the Europarl corpus.

2 Phrase-based Statistical Machine Transduction

The translation units in a phrase-based statistical translation system are bilingual phrases rather than simple paired words. Several systems that follow this approach have been presented in recent works (Zens et al., 2002; Koehn, 2004). These systems have demonstrated excellent translation performance in real tasks.

The basic idea of a phrase-based statistical machine translation system consists of the following steps (Zens et al., 2002): first, the source sentence is segmented into phrases; second, each source phrase is translated into a target phrase; and third, the target phrases are reordered in order to compose the target sentence.

Bilingual translation phrases are an important component of a phrase-based system. Different methods have been defined to obtain bilingual translations phrases, mainly from word-based alignments and from syntax-based models (Yamada and Knight, 2001).

In this work, we focus on learning bilingual word phrases by using Stochastic Inversion Transduction Grammars (SITGs) (Wu, 1997). This formalism al-

Proceedings of the Workshop on Statistical Machine Translation, pages 130–133,
New York City, June 2006.

lows us to obtain bilingual word phrases in a natural way from the bilingual parsing of two sentences. In addition, the SITGs allow us to easily incorporate many desirable characteristics to word phrases such as length restrictions, selection according to the word alignment probability, bracketing information, etc. We review this formalism in the following section.

3 Stochastic Inversion Transduction Grammars

Stochastic Inversion Transduction Grammars (SITGs) (Wu, 1997) can be viewed as a restricted subset of Stochastic Syntax-Directed Transduction Grammars. They can be used to simultaneously parse two strings, both the source and the target sentences. SITGs are closely related to Stochastic Context-Free Grammars.

Formally, a SITG in Chomsky Normal Form[1] τ_s can be defined as a tuple (N, S, W_1, W_2, R, p), where: N is a finite set of non-terminal symbols; $S \in N$ is the axiom of the SITG; W_1 is a finite set of terminal symbols of language 1; and W_2 is a finite set of terminal symbols of language 2. R is a finite set of: lexical rules of the type $A \rightarrow x/\epsilon$, $A \rightarrow \epsilon/y$, $A \rightarrow x/y$; direct syntactic rules that are noted as $A \rightarrow [BC]$; and inverse syntactic rules that are noted as $A \rightarrow \langle BC \rangle$, where $A, B, C \in N$, $x \in W_1$, $y \in W_2$, and ϵ is the empty string. When a direct syntactic rule is used in a parsing, both strings are parsed with the syntactic rule $A \rightarrow BC$. When an inverse rule is used in a parsing, one string is parsed with the syntactic rule $A \rightarrow BC$, and the other string is parsed with the syntactic rule $A \rightarrow CB$. Term p of the tuple is a function that attaches a probability to each rule.

An efficient Viterbi-like parsing algorithm that is based on a Dynamic Programing Scheme is proposed in (Wu, 1997). The proposed algorithm has a time complexity of $O(|x|^3|y|^3|R|)$. It is important to note that this time complexity restricts the use of the algorithm to real tasks with short strings.

If a bracketed corpus is available, then a modified version of the parsing algorithm can be defined to take into account the bracketing of the strings. The modifications are similar to those proposed in (Pereira and Schabes, 1992) for the *inside* algorithm. Following the notation that is presented in (Pereira and Schabes, 1992), we can define a partially bracketed corpus as a set of sentence pairs that are annotated with parentheses that mark constituent frontiers. More precisely, a bracketed corpus Ω is a set of tuples (x, B_x, y, B_y), where x and y are strings, B_x is the bracketing of x, and B_y is the bracketing of y. Let d_{xy} be a parsing of x and y with the SITG τ_s. If the SITG does not have useless symbols, then each non-terminal that appears in each sentential form of the derivation d_{xy} generates a pair of substrings $x_i \ldots x_j$ of x, $1 \leq i \leq j \leq |x|$, and $y_k \ldots y_l$ of y, $1 \leq k \leq l \leq |y|$, and defines a *span* (i, j) of x and a *span* (k, l) of y. A derivation of x and y is compatible with B_x and B_y if all the spans defined by it are compatible with B_x and B_y. This compatibility can be easily defined by the function $c(i, j, k, l)$, which takes a value of 1 if (i, j) does not overlap any $b \in B_x$ and, if (k, l) does not overlap any $b \in B_y$; otherwise it takes a value of 0. This function filters those derivations (or partial derivations) whose parsing is not compatible with the bracketing defined in the sample (Sánchez and Benedí, 2006).

The algorithm can be implemented to compute only those subproblems in the Dynamic Programming Scheme that are compatible with the bracketing. Thus, the time complexity is $O(|x|^3|y|^3|R|)$ for an unbracketed string, while the time complexity is $O(|x||y||R|)$ for a fully bracketed string. It is important to note that the last time complexity allows us to work with real tasks with longer strings.

Moreover, the parse tree can be efficiently obtained. Each node in the tree relates two word phrases of the strings being parsed. The related word phrases can be considered to be the translation of each other. These word phrases can be used to compute the translation table of a phrase-based machine statistical translation system.

4 Experiments

The experiments in this section were carried out for the shared task proposed in this workshop. This consisted of building a probabilistic phrase translation table for phrase-based statistical machine translation. Evaluation was translation quality on an unseen test set. The experiments were carried out using

[1] A Normal Form for SITGs can be defined (Wu, 1997) by analogy to the Chomsky Normal Form for Stochastic Context-Free Grammars.

the Europarl corpus (Koehn, 2005). Table 1 shows the language pairs and some figures of the training corpora. The test set had $3,064$ sentences.

Languages	Sentences	# words (input/output)
De-En	751,088	15,257,871 / 16,052,702
Es-En	730,740	15,725,136 / 15,222,505
Fr-En	688,031	15,599,184 / 13,808,505

Table 1: Figures of the training corpora. The languages are English (En), French (Fr), German (De) and Spanish (Es)

A common framework was provided to all the participants so that the results could be compared. The material provided comprised of: a training set, a language model, a baseline translation system (Koehn, 2004), and a word alignment. The participants could augment these items by using: their own training corpus, their own sentence alignment, their own language model, or their own decoder. We only used the provided material for the experiments reported in this work. The BLEU score was used to measure the results.

A SITG was obtained for every language pair in this section as described below. The SITG was used to parse paired sentences in the training sample by using the parsing algorithm described in Section 3. All pairs of word phrases that were derived from each internal node in the parse tree, except the root node, were considered for the phrase-based machine translation system. A translation table was obtained from paired word phrases by placing them in the adequate order and counting the number of times that each pair appeared in the phrases. These values were then appropriately normalized (Sánchez and Benedí, 2006).

4.1 Obtaining a SITG from an aligned corpus

For this experiment, a SITG was constructed for every language pair as follows. The alignment was used to compose lexical rules of the form $A \rightarrow e/f$. The probability of each rule was obtained by counting. Then, two additional rules of the form $A \rightarrow [AA]$ and $A \rightarrow \langle AA \rangle$ were added. It is important to point out that the constructed SITG did not parse all the training sentences. Therefore, the model was *smoothed* by adding all the rules of the form $A \rightarrow e/\epsilon$ and $A \rightarrow \epsilon/f$ with low probability, so that all the training sentences could be parsed. The rules were then adequately normalized.

This SITG was used to obtain word phrases from the training corpus. Then, these word phrases were used by the Pharaoh system (Koehn, 2004) to translate the test set. We used word phrases up to a given length. In these experiments several lengths were tested and the best values ranged from 6 to 10. Table shows 2 the obtained results and the size of the translation table.

Lang.	BLEU	Lang.	BLEU
De-En	15.91 (8.7)	En-De	11.20 (9.7)
Es-En	22.85 (6.5)	En-Es	21.18 (8.6)
Fr-En	21.30 (7.3)	En-Fr	20.12 (8.1)

Table 2: Obtained results for different pairs and directions. The value in parentheses is the number of word phrases in the translation table (in millions).

Note that better results were obtained when English was the target language.

4.2 Using bracketing information in the parsing

As Section 3 describes, the parsing algorithm for SITGs can be adequately modified in order to take bracketed sentences into account. If the bracketing respects linguistically motivated structures, then aligned phrases with linguistic information can be used. Note that this approach requires having quality parsed corpora available. This problem can be reduced by using automatically learned parsers.

This experiment was carried out to determine the performance of the translation when some kind of structural information was incorporated in the parsing algorithm described in Section 3. We bracketed the English sentences of the Europarl corpus with an automatically learned parser. This automatically learned parser was trained with bracketed strings obtained from the UPenn Treebank corpus. We then obtained word phrases according to the bracketing by using the same SITG that was described in the previous section. The obtained phrases were used with the Pharaoh system. Table 3 shows the results obtained in this experiment.

Note that the results decreased slightly in all

Lang.	BLEU	Lang.	BLEU
De-En	15.13 (7.1)	En-De	10.40 (9.2)
Es-En	21.61 (6.6)	En-Es	19.86 (9.6)
Fr-En	20.57 (6.3)	En-Fr	18.95 (8.3)

Table 3: Obtained results for different pairs and directions when word phrases were obtained from a parsed corpus.The value in parentheses is the number of word phrases in the translation table (in millions).

cases. This may be due to the fact that the bracketing incorporated hard restrictions to the paired word phrases and some of them were too forced. In addition, many sentences could not be parsed (up to 5% on average) due to the bracketing. However, it is important to point out that incorporating bracketing information to the English sentences notably accelerated the parsing algorithm, thereby accelerating the process of obtaining word phrases, which is an important detail given the magnitude of this corpus.

4.3 Combining word phrases

Finally, we considered the combination of both kinds of segments. The results can be seen in Table 4. This table shows that the results improved the results of Table 2 when English was the target language. However, the results did not improve when English was the source language. The reason for this could be that both kinds of segments were different in nature, and, therefore, the number of word phrases increased notably, specially in the English part.

Lang.	BLEU	Lang.	BLEU
De-En	16.39 (17.1)	En-De	11.02 (15.3)
Es-En	22.96 (11.7)	En-Es	20.86 (14.1)
Fr-En	21.73 (17.0)	En-Fr	19.93 (14.9)

Table 4: Obtained results for different pairs and directions when word phrases were obtained from a non-parsed corpus and a parsed corpus.The value in parentheses is the number of word phrases in the translation table (in millions).

5 Conclusions

In this work, we have explored the problem of obtaining word phrases for phrase-based machine translation systems from SITGs. We have described how the parsing algorithms for this formalism can be modified in order to take into account a bracketed corpus. If bracketed corpora are used the time complexity can decrease notably and large tasks can be considered. Experiments were reported for the Europarl corpus, and the results obtained were competitive.

For future work, we propose to work along different lines: first, to incorporate new linguistic information in both the parsing algorithm and in the aligned corpus; second, to obtain better SITGs from aligned bilingual corpora; an third, to improve the SITG by estimating the syntactic rules. We also intend to address other machine translation tasks.

Acknowledgements

This work has been partially supported by the *Universidad Politécnica de Valencia* with the ILETA project.

References

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. of AMTA*.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit*.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

F. Pereira and Y. Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135. University of Delaware.

J.A. Sánchez and J.M. Benedí. 2006. 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*, page Accepted, Oslo, Norway.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proc. of the 39th Annual Meeting of the Association of Computational Linguistics*, pages 523–530.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In *Proc. of the 25th Annual German Conference on Artificial Intelligence*, pages 18–32.