Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation

David Vilar, Daniel Stein and Hermann Ney

Department for Computer Science 6
RWTH Aachen University, Germany
{vilar,stein,ney}@i6.informatik.rwth-aachen.de

Abstract

Similar to phrase-based machine translation, hierarchical systems produce a large proportion of phrases, most of which are supposedly junk and useless for the actual translation. For the hierarchical case, however, the amount of extracted rules is an order of magnitude bigger. In this paper, we investigate several soft constraints in the extraction of hierarchical phrases and whether these help as additional scores in the decoding to prune unneeded phrases. We show the methods that help best.

1. Introduction

Translation based on word groups (so called phrases) has proven to be an effective approach to statistical machine translation, and a great improvement over the initial word-based approaches first presented in the seminal IBM paper [1]. However these models are still limited. They do not consider long range dependencies, as the phrases must be contiguous both in the source and the target language, and the reordering of the phrases is not directly modelled and is normally applied in a heuristic way.

The hierarchical phrase based translation model [2] addresses these issues by allowing "gaps" in the phrases. In this way the reorderings are integrated in the translation process, and non-contiguous word groups can be translated in a consistent way.

In this work we explore various refinements to the phrase extraction algorithm and the effect on the final translation quality. Additionally, we add syntax information to the extracted phrases, computing features which measure how well the extracted phrases correspond to linguistic structures. Since the hierarchical phrase extraction process allows for a high number of phrases, most of them of dubious quality for the translation, we expect this approach to be useful as an additional knowledge source for the decoding process.

The advantages of our approach in comparison to other recent syntax-based approaches is on the one hand its simplicity and thus flexibility, which allows it also to be integrated with standard phrase-based approaches, and on the other hand the possibility to use syntax information both on the source and target languages.

Results on the recent IWSLT 2008 evaluation are reported, where our hierarchical system resulted in improvements over our state-of-the-art phrase-based translation system. The paper is organized as follows: in Section 2 we analyze related work on the same area. In Section 3 we describe our baseline system. Section 4 and Section 5 constitute the main contribution in terms of refinements to the extraction process and including syntax information. Section 6 presents experimental results, which are analyzed in Section 7. Section 8 concludes the paper.

2. Related Work

The hierarchical phrase based approach was first presented by David Chiang in [3] and further detailed in [2]. Already in [3], Chiang proposes the use of syntactic information together with his new hierarchical approach, but without success.

Some recent publications have shown that the use of syntax for translation achieves significant improvements. One prominent example are the works by the ISI group (e.g. [4, 5]). These works go apart from the standard phrase-based approach by defining new translation units and extraction procedures, but they try to still keep the advantages of phrase-based translation [6].

There have also been some previous efforts in combining syntactic information together with a hierarchical phrase-based approach, see for example Zollmann

中 $X^{\sim 0}$ 那个 $X^{\sim 1}$ # It's the $X^{\sim 1}$ in the $X^{\sim 0}$
也 要 $X^{\sim 0}$ 一些 $X^{\sim 1}$ # like to $X^{\sim 0}$ some $X^{\sim 1}$ too

Figure 1: Example of hierarchical rules.

et al. [7] or more recently Marton et al. [8].

Our work differs from the above mentioned mainly in that we extract the syntactic information already at the training phrase, and it gets integrated in the search process as an additional model in the base log-linear combination that underlies most state-of-the-art statistical machine translation systems. Therefore, no modification of the search algorithms is needed and we can also make use of syntactic information for both languages, source and target (see also Section 5). Most of the previous work limit themselves only to the target language side, as the correspondences between the syntactic structures of both languages are hard to define.

3. Hierarchical Phrase Based Translation

The baseline system we use is an in-house implementation of a hierarchical phrase-based system, similar to the one presented in [2]. This approach can be seen as an extension of the phrase-based approach [9, 10], where we allow for "gaps" in the extracted phrases. In this way, longer range dependencies in the translation process can be modelled and the reordering is directly integrated in the decoding process.

The translation model can be formalized as a synchronous context free grammar, where the rules are of the form

$$X \rightarrow \langle \gamma, \alpha, \sim \rangle$$

where X is a non-terminal, γ and α are strings of terminals (respectively in the source and target languages) and non-terminals, and $\sim$ is a one-to-one correspondence between the non-terminals of α and γ, which shows corresponding "elided" parts in the source and target sentences. An example of rules can be seen in Figure 1. Additionally, so called "glue rules" are added which allow for the concatenation of translation of subparts of the source sentence in a monotonic way. For further details the reader is referred to [2].

The extraction process starts with the same phrase-extraction procedure as in the standard phrase-based translation, i.e. sequences of source and target words that are aligned only to each other. We then proceed to generate new rules as follows: for each of the extracted phrase pairs, smaller sub-phrases are sought for. If found, the corresponding parts are substituted by a non-terminal and linked together by the relation $\sim$. This process is then iterated with the new extracted rules, until the desired maximum number of non-terminals is achieved. Note however, that only standard phrases are made into gaps.

In his original work, Chiang proposes following constraints to the extraction process:

1. Rules may have a maximum of two non-terminals.
2. Non-terminals must be non-adjacent in the source side.
3. Rules must have at least one terminal symbol.
4. Some additional rule length constraints are applied for efficiency.
5. Only minimal initial phrases are included (i.e. initial phrases are not "expanded" over non-aligned words).

The phrases get scored by relative frequencies, whereas for the hierarchical phrases the counts of the originating standard phrases get distributed among all the generated hierarchical rules. These scores are computed for the translation directions source-to-target and target-to-source, which get combined log-lineally with additional IBM1-like word level scores at the phrase level, word and phrase penalty scores at generation time.

The decoding process is basically a parsing of the source sentence according to the defined grammar, keeping track of the target language translation contexts in order to compute language model scores during the translation process. In order to deal with the high computational effort of the search process, early pruning is carried out in the form of "cube pruning" or its lazy version "cube growing" [11].

4. Refinement of Extraction Heuristics

In standard phrase-based translation, the extraction of additional phrase pairs by including non-aligned words adjacent to the standard phrases, both in the source and target language, has proven to be beneficial in the translation process. However, in [2] they are not included in the extraction process, probably for efficiency reasons.

We however, found it beneficial to include them in the translation process.

Although difficult to justify from a practical point of view, practice has shown that simple heuristics often help in the translation process. For example, simply adding a count of how many words are generated (the so-called "word penalty") helps controlling the length of the produced translations and is now a standard component in most phrase-based translation systems. In this work we propose and evaluate some additional heuristic features which are straightforward to implement and include in the baseline system.

We tested the following features, the names in brackets represent the entry in the result tables:

Paste rule (=paste) We call *paste rules* those rules of the form

$$X \rightarrow \langle X^{\sim 0}\alpha, X^{\sim 0}\beta\rangle \text{ or } X \rightarrow \langle \alpha X^{\sim 0}, \beta X^{\sim 0}\rangle$$

We include a binary feature which is activated for each phrase of this form. These rules contrast with "reordering rules" and adjusting the weight of the corresponding scaling factor, we can control how much reordering we allow in the translation system.

Hierarchical penalty (=hierarch) A binary feature which indicates that we are using a hierarchical rule. Adjusting the weight of this feature we can give more weight to the hierarchical or to the standard phrases.

Number of non-terminals (=1NT2NT) Two binary features indicating if the rule has one or two non-terminals.

Extended glue rule (=glue2) We add a rule of the form

$$X \rightarrow \langle X^{\sim 0}X^{\sim 1}, X^{\sim 0}X^{\sim 1}\rangle$$

similar to the original glue rule proposed in [2] (the original glue rule has the starting non-terminal S on the left hand side). The inclusion of this rule allows to concatenate the translation of different phrases in a monotonic way also inside of a gap, not only at the top-most level.

5. Syntactical Features

In this section, we propose the inclusion of additional syntactically motivated features. In contrast to other approaches in which rules are extracted to try to enforce the syntactical integrity of the translation (e.g. [4]), we do not limit the extraction algorithm. The rule extraction algorithm is left untouched as presented in Section 3, and additional scores are computed for the generated phrases. In this way, by adjusting the corresponding scaling factor we can fall back to the original system.

The inclusion of the syntactic information at training time has also additional advantages. In contrast to other approaches, which normally only consider the target syntax, we can analyze both the source and the target part of the rules and thus the system is able to make a better usage of bilingual correspondences between syntactic structures. Furthermore, the inclusion of this information as additional scores in the phrases does not have any impact on computation time.

Our goal is to determine if the bilingual phrase pair corresponds to some syntactic structures, or not. We stress again that we do not limit the amount of phrases we extract, as non-syntactical phrases are necessary to achieve good translation performance [6]. We parse the English part with Charniak's parser[1] and the Chinese part with the Stanford parser [12].

Given an initial phrase pair we analyze both parts (source and target) independently. In order to weight a phrase ranging from position i to j, we check whether this sequence corresponds to the yield of some node in the parse tree. If not, we first determine the minimum number of words that we have to delete or add to the phrase so that it can be associated with such a node. In order to compute this number we search in the tree in a bottom-up manner, looking for the lowest node that does not cover all words, or in a top-down manner, looking for the highest node that covers all the words of the phrase and possibly more. From this we can compute the desired quantity, which we denote with $m(i, j)$.

If the phrase matches a node completely, it always gets a count of 1. If not, we propose different variants for the scoring of the phrases:

- A count of zero. In this way we define a binary feature in which we strictly say if the phrase corresponds to a syntactic structure.
- One divided by $m(i, j)$. In this way we can capture the notion of being "quasi-syntactically cor-

[1] `http://www.cs.brown.edu/people/ec`

rect", in the sense that if only few words are missing (or are too much) we do not penalize the phrase so drastically.

- One divided by the exponential of $m(i,j)$. This is similar to the preceding one but with an exponential decay.

- The fraction of "correct" words in the phrase. In this way the size of the phrase is also taken into account.

These different possibilities are summarized as follows:

$$c(i,j) = \begin{cases} \delta\left(m(i,j),0\right) & \text{Binary} \\ \dfrac{1}{m(i,j)+1} & \text{Linear} \\ \dfrac{1}{\exp\left(m(i,j)\right)} & \text{Exponentional} \\ \dfrac{j-i}{(j-i)+m(i,j)} & \text{Relative} \end{cases}$$

These counts are then added up for every occurrence of a phrase pair in the training data and then normalized with the total count of the given phrase pair. In this way we get a relative frequency for the syntax component of the phrase inventory.

For hierarchical phrases, we add all counts from the phrases involved (that is, the overall initial phrase plus every initial phrase that is replaced with a non-terminal) and divide this by the number of phrases, in order to maintain a normalized value. In this way, the syntactic well-formedness of a hierarchical phrase is dependent of the original phrase from which it originates and the gaps that we introduce.

In Figure 2, an example for a hierarchical phrase is presented. Both the word spans of "the public toilet" and "Where is the public toilet" completely match the yields of nodes in the parse tree (S and NP respectively). Thus, they would both get a linguistic count of 1, and likewise the hierarchical phrase "Where is $X^{\sim 0}$". The phrase "is the public", however, does not match any particular node, and has one word missing to complete the next higher node (VP). Thus, it would get a 0 for the binary count, 0.5 for the linear count, $1/2.718 = 0.36$ for log count and $4/5 = 0.8$ as relative count.

6. Experiments

Experimental results are presented on the BTEC task of the IWSLT 2008 Evaluation. The BTEC corpus ("Basic Travel Expressions Corpus") is a bilingual corpus in the touristic domain and is used regularly in the IWSLT evaluations that take place on a yearly basis. This year's evaluation included several conditions an language pairs: Chinese-to-English, Arabic-to-English, Chinese-to-Spanish, English-to-Chinese and Chinese-to-English-to-Spanish, most of them with text and speech conditions. In this work we concentrate on the Chinese-to-English language pair, for the text condition.

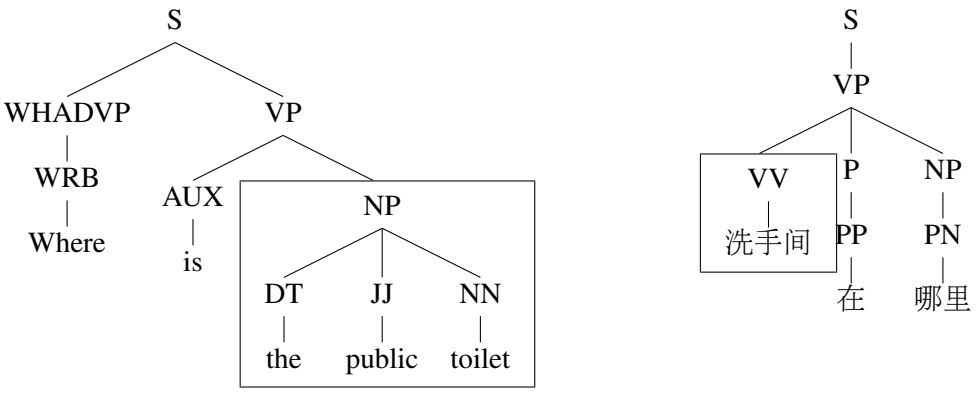
The statistics of the corpora can be found in Table 1. We selected the test corpus of the 2004 evaluation campaign as development set on which we optimize the weights of the log-linear model combination maximizing the BLEU score, in a manner similar to [13]. The test corpus of the 2005 evaluation campaign was used as blind test set for system selection and system combination optimization. The rest of the provided development corpora were added to the training data. For the English part of the added development data we included the two longest references, following [14]. The training corpus with the development data included is denoted "Extended training data" in Table 1.

The Chinese side of the corpus was preprocessed using the ICTClass word segmenter [15]. For the English part we performed tokenization of punctuation marks and contraction expansion. No further preprocessing was applied to the data.

The results are presented in Table 2. It can be seen from this table that the performance of the hierarchical system is improved by incorporating additional information, both in the form of the simple heuristics presented in Section 4 and syntactic information introduced in Section 5. It is however somewhat difficult to further interpret the results and give clear conclusions as to which method works best, although it seems clear that the syntax information outperforms the other methods.

Concerning the heuristic scores, each feature separately improves the translation quality by 0.2-0.5% BLEU on the blind test set (test05). The combination of certain features seems to improve the performance but some other subsets actually degrade the BLEU score.

The different syntactic features seem to perform equally good, obtaining an improvement of 0.7%



$X^{\sim 0}$ 在 哪里 # Where is $X^{\sim 0}$

Figure 2: Example of a syntax-enhanced hierarchical rule.

		Chinese	English
Training data	Sentences	19 972	
	Running words	171 932	188 704
	Vocabulary	8 535	7 567
Extended training data	Sentences	23 940	
	Running words	181 486	232 746
	Vocabulary	9 041	10 350
Test 2004 Data	Sentences	500	
	Running words	7 543	10 718
	OOVs	96	154
Test 2005 Data	Sentences	506	
	Running words	8 052	10 828
	OOVs	101	164
Test 2008 Data	Sentences	507	
	Running words	6325	
	OOVs	87	

Table 1: Corpus statistics of the IWSLT BTEC Data. “Extended training data” denotes the training set together with the 2003, 2006 and 2007 development and test datasets.

BLEU, although the exponential decay function seems to perform better in terms of TER. The inclusion of the heuristic features together with the syntax information improves the translation performance in some cases.

7. Discussion

We believe that the results presented in the previous section must be interpreted cautiously. Table 2 also shows the performance of the different systems on the test04 data which we used as development set for parameter tuning. In this corpus the variability of the results is much greater and no clear conclusions can be drawn. This can partly be due to the relatively small size of the data used and also due to instabilities in the optimization algorithm used (downhill simplex), as pointed out in Lambert et al. [16]. The consistent results obtained on the blind test data set, however, give a reasonable indication that the proposed methods improve translation quality.

8. Conclusions

We have analyzed the effect on translation quality of different extraction heuristics for a hierarchical phrase-based translation system. We also have shown how to include syntactic information at the training phase. In this way, we can include syntactic information from both languages at translation time without the need of modification on a decoder. Experiments were reported on the IWSLT 2008 evaluation task, which, although with a high degree of variability, shows the adequateness of our methods.

9. Acknowledgements

We would like to thank Niklas Hoppe for his help in conducting the experiments.

This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) under the project "Statistische Text´übersetzung" (NE 572/5q).

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

10. 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*, vol. 19, no. 2, pp. 263–311, June 1993.

[2] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[3] ——, "A Hierarchical Phrase-Based Model for Statistical Machine Translation," in *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 263–270. [Online]. Available: http://www.aclweb.org/anthology/P/P05/P05-1033

[4] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a translation rule," *Proceedings of HLT/NAACL*, vol. 4, pp. 273–280, 2004.

[5] D. Marcu, W. Wang, A. Echihabi, and K. Knight, "SPMT: Statistical machine translation with syntactified target language phrases," *Proceedings of EMNLP*, pp. 44–52, 2006.

[6] S. DeNeefe, K. Knight, W. Wang, and D. Marcu, "What Can Syntax-based MT Learn from Phrase-based MT?" *Proc. EMNLP/CONLL*, 2007.

[7] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," *Proc. of the Workshop on Statistical Machine Translation, HLT/NAACL, New York, June*, 2006.

[8] Y. Marton and P. Resnik, "Soft Syntactic Constraints for Hierarchical Phrased-Based Translation," *Proceedings of ACL-08: HLT*, pp. 1003–1011, June 2008.

[9] R. Zens, F. J. Och, and H. Ney, "Phrase-based statistical machine translation," in *25th German Conf. on Artificial Intelligence (KI2002)*, ser. Lecture Notes in Artificial Intelligence (LNAI), M. Jarke, J. Koehler, and G. Lakemeyer, Eds., vol. 2479. Aachen, Germany: Springer Verlag, September 2002, pp. 18–32.

	test04		test05		test08
	BLEU	TER	BLEU	TER	BLEU
baseline	47.3	42.6	50.9	37.6	39.6
non-syntactic information					
hierarch	48.4	41.9	51.4	38.1	39.6
paste	49.1	41.6	51.1	38.0	40.8
glue2	48.2	41.8	51.2	37.6	39.7
1NT2NT	48.4	42.2	51.8	37.2	39.8
hierarch + paste	48.5	42.0	51.9	37.6	39.6
hierarch + paste + glue2	49.2	42.5	50.8	37.5	39.5
hierarch + paste + glue2 + 1NT2NT	48.6	41.6	51.0	37.9	40.0
syntactic information					
binary	47.8	41.7	51.7	37.5	40.3
linear	47.6	41.9	51.2	37.6	40.6
exponential	47.9	41.7	51.6	37.4	40.3
relative	47.3	42.4	51.5	37.3	40.2
combination of both syntactic and non-syntactic information (all features)					
binary	46.9	42.5	50.6	38.4	39.9
linear	48.0	42.3	51.2	38.0	40.5
exponential	47.7	42.3	51.0	38.4	40.0
relative	47.8	42.3	51.0	38.0	40.3

Table 2: Results on the IWSLT data. The IWSLT 2008 evaluation server does not provide TER scores.

reference	Where is the exchange counter ?
baseline	The currency exchange office is
syntactical	Where is the currency exchange office ?

reference	Could you exchange it for a new one ?
baseline	You can buy a new one ?
syntactical	Could you change it for a new one ?

reference	You can take our airport shuttle bus to pick up the car .
baseline	You can take our airport shuttle bus with me .
syntax	You can take our the airport shuttle bus come to pick it up .

Figure 3: Sample sentences with changes due to syntactical feature

[10] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, Edmonton, Canada, May/June 2003, pp. 127–133.

[11] L. Huang and D. Chiang, "Forest Rescoring: Faster Decoding with Integrated Language Models," *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 144–151, June 2007.

[12] D. Klein and C. Manning, "Accurate unlexicalized parsing," *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

[13] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[14] A. Mauser, D. Vilar, G. Leusch, Y. Zhang, and H. Ney, "The RWTH Machine Translation System for IWSLT 2007," in *International Workshop on Spoken Language Translation*, Trento, Italy, Oct. 2007, pp. 161–168.

[15] H. Zhang, Q. Liu, X. Cheng, H. Zhang, and H. Yu, "Chinese Lexical Analysis Using Hierarchical Hidden Markov Model," *Proceedings of the Second SIGHAN Workshop*, pp. 63–70, 2003.

[16] P. Lambert and R. Banchs, "Tuning Machine Translation Parameters with SPSA," *Proceedings of IWSLT*, pp. 190–196, 2006.