Simple and Effective Approach for Consistent Training of Hierarchical Phrase-based Translation Models

Stephan Peitz[1] and **David Vilar**[2] and **Hermann Ney**[1]

[1] Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{peitz,ney}@cs.rwth-aachen.de

[2] Pixformance GmbH
D-10587 Berlin, Germany
david.vilar@gmail.com

Abstract

In this paper, we present a simple approach for consistent training of hierarchical phrase-based translation models. In order to consistently train a translation model, we perform hierarchical phrase-based decoding on training data to find derivations between the source and target sentences. This is done by synchronous parsing the given sentence pairs. After extracting k-best derivations, we reestimate the translation model probabilities based on collected rule counts. We show the effectiveness of our procedure on the IWSLT German→English and English→French translation tasks. Our results show improvements of up to 1.6 points BLEU.

1 Introduction

In state of the art statistical machine translation systems, the translation model is estimated by following heuristic: Given bilingual training data, a word alignment is trained with tools such as GIZA++ (Och and Ney, 2003) or *fast_align* (Dyer et al., 2013). Then, all valid translation pairs are extracted and the translation probabilities are computed as relative frequencies (Koehn et al., 2003).

However, this extraction method causes several problems. First, this approach does not consider, whether a translation pair is extracted from a likely alignment or not. Further, during the extraction process, models employed in decoding are not considered.

For phrase-based translation, a successful approach addressing these issues is presented in (Wuebker et al., 2010). By applying a phrase-based decoder on the source sentences of the training data and constraining the translations to the corresponding target sentences, k-best segmentations are produced. Then, the phrases used for these segmentations are extracted and counted. Based on the counts, the translation model probabilities are recomputed. To avoid over-fitting, leave-one-out is applied.

However, for hierarchical phrase-based translation an equivalent approach is still missing.

In this paper, we present a simple and effective approach for consistent reestimation of the translation model probabilities in a hierarchical phrase-based translation setup. Using a heuristically extracted translation model as starting point, the training data are parsed bilingually. From the resulting hypergraphs, we extract k-best derivations and the rules applied in each derivation. This is done with a top-down k-best parsing algorithm. Finally, the translation model probabilities are recomputed based on the counts of the extracted rules. In our procedure, we employ leave-one-out to avoid over-fitting. Further, we consider all models which are used in translation to ensure a consistent training.

Experimental results are presented on the German→English and English→French IWSLT shared machine translation task (Cettolo et al., 2013). We are able to gain improvements of up to 1.6% BLEU absolute and 1.4% TER over a competitive baseline. On all tasks and test sets, the improvements are statistically significant with at least 99% confidence.

The paper is structured as follow. First, we revise the state of the art hierarchical phrase-based extraction and translation process. In Section 3, we propose our training procedure. Finally, experimental results are given in Section 4 and we conclude with Section 5.

2 Hierarchical Phrase-based Translation

In hierarchical phrase-based translation (Chiang, 2005), discontinuous phrases with “gaps” are allowed. The translation model is formalized as a synchronous context-free grammar (SCFG) and consists of bilingual rules, which are based on bilingual standard phrases and discontinuous phrases. Each bilingual rule rewrites a generic non-terminal X into a pair of strings $\tilde{f}$ and $\tilde{e}$ with both terminals and non-terminals in both languages

$$X \rightarrow \langle \tilde{f}, \tilde{e} \rangle. \tag{1}$$

In a standard hierarchical phrase-based translation setup, obtaining these rules is based on a heuristic extraction from automatically word-aligned bilingual training data. Just like in the phrase-based approach, all bilingual rules of a sentence pair are extracted given an alignment. The standard phrases are stored as *lexical rules* in the rule set. In addition, whenever a phrase contains a sub-phrase, this sub-phrase is replaced by a generic non-terminal X. With these hierarchical phrases we can define the *hierarchical rules* in the SCFG. The rule probabilities which are in general defined as relative frequencies are computed based on the joint counts $C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)$ of a bilingual rule $X \rightarrow \langle \tilde{f}, \tilde{e} \rangle$

$$p_H(\tilde{f}|\tilde{e}) = \frac{C(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)}. \tag{2}$$

The translation probabilities are computed in source-to-target as well as in target-to-source direction. In the translation processes, these probabilities are integrated in the log-linear combination among other models such as a language model, word lexicon models, word and phrase penalty and binary features marking hierarchical phrases, glue rule and rules with non-terminals at the boundaries.

The translation process of hierarchical phrase-based approach can be considered as parsing problem. Given an input sentence in the source language, this sentence is parsed using the source language part of the SCFG. In this work, we perform this step with a modified version of the CYK+ algorithm (Chappelier and Rajman, 1998). The output of this algorithm is a *hypergraph*, which represents all possible *derivations* of the input sentence. A derivation represents an application of rules from the grammar to generate the given input sentence. Using the the associated target part of the applied rule, for each derivation a translation can be constructed. In a second step, the language model score is incorporated. Given the hypergraph, this is done with the cube pruning algorithm presented in (Chiang, 2007).

3 Translation Model Training

We propose following pipeline for consistent hierarchical phrase-based training: First we train a word alignment, from which the baseline translation model is extracted as described in the previous section. The log-linear parameter weights are tuned with MERT (Och, 2003) on a development set to produce the baseline system. Next, we perform decoding on the training data. As the translations are constrained to the given target sentences, we name this step *forced decoding* in the following. Details are given in the next subsection. Given the counts $C_{FD}(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)$ of the rules, which have been applied in the forced decoding step, the translation probabilities $p_{FD}(\tilde{f}|\tilde{e})$ for the translation model are recomputed:

$$p_{FD}(\tilde{f}|\tilde{e}) = \frac{C_{FD}(X \rightarrow \langle \tilde{f}, \tilde{e} \rangle)}{\sum_{\tilde{f}'} C_{FD}(X \rightarrow \langle \tilde{f}', \tilde{e} \rangle)}. \tag{3}$$

Finally, using the translation model with the reestimated probabilities, we retune the log-linear parameter weights and obtain our final system.

3.1 Forced Decoding

In this section, we describe the forced decoding for hierarchical phrase-based translation in detail.

Given a sentence pair of the training data, we constrain the translation of the source sentence to produce the corresponding target sentence. For this constrained decoding process, the language model score is constant as the translation is fixed. Hence, the incorporation of the a language model is not needed. This results in a simplification of the decoding process as we do not have to employ the cube pruning algorithm as described in the previous section. Consequently, forced decoding for hierarchical phrase-based translation is equivalent to synchronous parsing of the training data. Dyer (2010) has described an approach to reduce the average-case run-time of synchronous parsing by splitting one bilingual parse into two successive monolingual parses. We adopt this method and first parse the source sentence and then the target sentence with CYK+.

If the given sentence pair has been parsed successfully, we employ a top-down k-best parsing algorithm (Chiang and Huang, 2005) on the resulting hypergraph to find the k-best derivations between the given source and target sentence. In this step, all models of the translation process are included (except for the language model). Further, leave-one-out is applied to counteract overfitting. Note, that the model weights of the baseline system are used to perform forced decoding.

Finally, we extract and count the rules which have been applied in the derivations. These counts are used to recompute the translation probabilities.

3.2 Recombination

In standard hierarchical phrase-based decoding, partial derivations that are indistinguishable from each other are recombined. In (Huck et al., 2013) two schemes are presented. Either derivations that produce identical translations or derivations with identical language model context are recombined. As in forced decoding the translation is fixed and a language model is missing, both schemes are not suitable.

However, a recombination scheme is necessary to avoid derivations with the same application of rules. Further, recombining such derivations increases simultaneously the amounts of considered derivations during k-best parsing. Given two derivations with the same set of applied rules, the order of application of the rules may be different. Thus, we propose following scheme for recombining derivations in forced decoding: Derivations that produce identical sets of applied rules are recombined. Figure 1 shows an example for $k = 3$. Employing the proposed scheme, derivations d_1 and d_2 are recombined since both share the same set of applied rules ($\{r_1, r_3, r_2\}$).

$$\begin{array}{llll}
d_1: & \{r_1, r_3, r_2\} & d_1: & \{r_1, r_3, r_2\} \\
d_2: & \{r_3, r_2, r_1\} & d_3: & \{r_4, r_5, r_1, r_2\} \\
d_3: & \{r_4, r_5, r_1, r_2\} & d_4: & \{r_6, r_5, r_2, r_3\} \\
 & \text{(a)} & & \text{(b)}
\end{array}$$

Figure 1: Example search space before (a) and after (b) applying recombination.

4 Experiments

4.1 Setup

The experiments were carried out on the IWSLT 2013 German→English shared translation task.[1] It is focusing the translation of TED talks. Bilingual data statistics are given in Table 1. The baseline system was trained on all available bilingual data and used a 4-gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998), trained with the SRILM toolkit (Stolcke, 2002). As additional data sources for the LM we selected parts of the Shuffled News and LDC English Gigaword corpora based on cross-entropy difference (Moore and Lewis, 2010). In all experiments, the hierarchical search was performed as described in Section 2.

	German	**English**	**English**	**French**
Sentences	4.32M		5.23M	
Run. Words	108M	109M	133M	147M
Vocabulary	836K	792K	845K	888K

Table 1: Statistics for the bilingual training data of the IWSLT 2013 German→English and English→French task.

To confirm the efficacy of our approach, additional experiments were run on the IWSLT 2013 English→French task. Statistics are given in Table 1.

The training pipeline was set up as described in the previous section. Tuning of the log-linear parameter weights was done with MERT on a provided development set. As optimization criterion we used BLEU (Papineni et al., 2001).

Forced decoding was performed on the TED talks portion of the training data (∼140K sentences). In both tasks, around 5% of the sentences could not be parsed. In this work, we just skipped those sentences.

We report results in BLEU [%] and TER [%] (Snover et al., 2006). All reported results are averages over three independent MERT runs, and we evaluated statistical significance with *MultEval* (Clark et al., 2011).

4.2 Results

Figure 2 shows the performance of setups using translation models with reestimated translation probabilities. The setups vary in the k-best derivation size extracted in the forced decoding (fd) step. Based on the performance on the development set, we selected two setups with $k = 500$ using leave-one-out (+l1o) and $k = 750$ without leave-one-out (-l1o). Table 2 shows the final results for the German→English task. Performing consistent translation model training improves the translation

[1] http://www.iwslt2013.org

	dev*		eval11		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
baseline	33.1	46.8	35.7	44.1	30.5	49.7
forced decoding -l1o	33.2	**46.3**	**36.3**	**43.4**	**31.2**	**48.8**
forced decoding +l1o	**33.6**	**46.2**	**36.6**	**43.0**	**31.8**	**48.3**

Table 2: Results for the IWSLT 2013 German→English task. The development set used for MERT is marked with an asterisk (*). Statistically significant improvements with at least 99% confidence over the baseline are printed in boldface.

	dev*		eval11		test	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]	BLEU [%]	TER [%]
baseline	28.1	55.7	37.5	42.7	31.7	49.5
forced decoding +l1o	**28.8**	**55.0**	**39.1**	**41.6**	**32.4**	**49.0**

Table 3: Results for the IWSLT 2013 English→French task. The development set used for MERT is marked with an asterisk (*). Statistically significant improvements with at least 99% confidence over the baseline are printed in boldface.

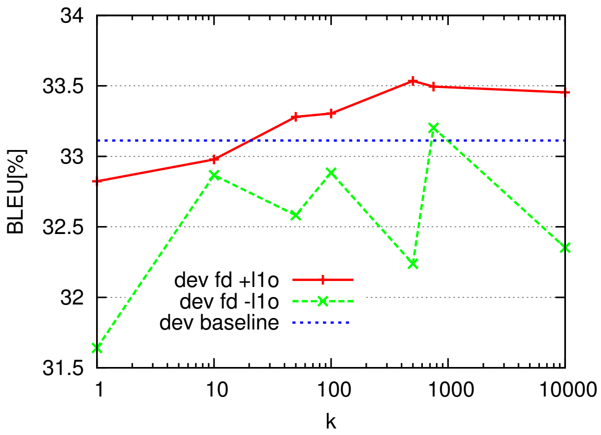

Figure 2: BLEU scores on the IWSLT German→English task of setups using translation models trained with different k-best derivation sizes. Results are reported on dev with (+l1o) and without leave-one-out (-l1o).

quality on all test sets significantly. We gain an improvement of up to 0.7 points in BLEU and 0.9 points in TER. Applying leave-one-out results in an additional improvement by up to 0.4 % BLEU and 0.5 % TER. The results for English→French are given in Table 3. We observe a similar improvement by up to 1.6 % BLEU and 1.1 % TER.

The improvements could be the effect of domain adaptation since we performed forced decoding on the TED talks portion of the training data. Thus, rules which were applied to decode the in-domain data might get higher translation probabilities.

Furthermore, employing leave-one-out seems to avoid overfitting as the average source rule length in training is reduced from 5.0 to 3.5 ($k = 500$).

5 Conclusion

We have presented a simple and effective approach for consistent training of hierarchical phrase-based translation models. By reducing hierarchical decoding on parallel training data to synchronous parsing, we were able to reestimate the translation probabilities including all models applied during the translation process. On the IWSLT German→English and English→French tasks, the final results show statistically significant improvements of up to 1.6 points in BLEU and 1.4 points in TER.

Our implementation was released as part of Jane (Vilar et al., 2010; Vilar et al., 2012; Huck et al., 2012; Freitag et al., 2014), the RWTH Aachen University open source statistical machine translation toolkit.[2]

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements no 287658 and no 287755.

[2] http://www.hltpr.rwth-aachen.de/jane/

References

Mauro Cettolo, Jan Nieheus, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, Heidelberg, Germany, December.

J.-C. Chappelier and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137, April.

Stanley F. Chen and Joshuo Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, August.

David Chiang and Liang Huang. 2005. Better k-best Parsing. In *Proceedings of the 9th Internation Workshop on Parsing Technologies*, pages 53–64, October.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, Michigan, June.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 176–181, Portland, Oregon, June.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT*, pages 644–648, Atlanta, Georgia, June.

Chris Dyer. 2010. Two monolingual parses are better than one (synchronous parse). In *In Proc. of HLT-NAACL*.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, April. To appear.

Matthias Huck, Jan-Thorsten Peter, Markus Freitag, Stephan Peitz, and Hermann Ney. 2012. Hierarchical Phrase-Based Translation with Jane 2. *The Prague Bulletin of Mathematical Linguistics*, 98:37–50, October.

Matthias Huck, David Vilar, Markus Freitag, and Hermann Ney. 2013. A performance study of cube pruning for large-scale hierarchical machine translation. In *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 29–38, Atlanta, Georgia, USA, June.

Reinerd Kneser and Hermann Ney. 1995. Improved backing-off for M-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processingw*, volume 1, pages 181–184, May.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta.

R.C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pages 220–224, Uppsala, Sweden, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

Andreas Stolcke. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open source hierarchical translation, extended with reordering and lexicon models. In *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pages 262–270, Uppsala, Sweden, July.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2012. Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, 26(3):197–216, September.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training phrase translation models with leaving-one-out. In *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, pages 475–484, Uppsala, Sweden, July.