Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation

Matthias Huck[1] and **David Vilar**[1,2] and **Daniel Stein**[1] and **Hermann Ney**[1]

[1] Human Language Technology and Pattern Recognition Group, RWTH Aachen University
<surname>@cs.rwth-aachen.de

[2] DFKI GmbH
Berlin, Germany
david.vilar@dfki.de

Abstract

In this paper we apply lightly-supervised training to a hierarchical phrase-based statistical machine translation system. We employ bitexts that have been built by automatically translating large amounts of monolingual data as additional parallel training corpora. We explore different ways of using this additional data to improve our system.

Our results show that integrating a second translation model with only non-hierarchical phrases extracted from the automatically generated bitexts is a reasonable approach. The translation performance matches the result we achieve with a joint extraction on all training bitexts while the system is kept smaller due to a considerably lower overall number of phrases.

1 Introduction

We investigate the impact of an employment of large amounts of unsupervised parallel data as training data for a statistical machine translation (SMT) system. The unsupervised parallel data is created by automatically translating monolingual source language corpora. This approach is called *lightly-supervised training* in the literature and has been introduced by Schwenk (2008). In contrast to Schwenk, we do not apply lightly-supervised training to a conventional phrase-based system (Och et al., 1999; Koehn et al., 2003) but to a hierarchical phrase-based translation (HPBT) system.

In hierarchical phrase-based translation (Chiang, 2005) a weighted synchronous context-free grammar is induced from parallel text, the search is based on CYK+ parsing (Chappelier and Rajman, 1998) and typically carried out using the cube pruning algorithm (Huang and Chiang, 2007). In addition to the contiguous *lexical phrases* as in standard phrase-based translation, the hierarchical phrase-based paradigm also allows for phrases with gaps which are called *hierarchical phrases*. A generic non-terminal symbol serves as a placeholder that marks the gaps.

In this paper we study several different ways of incorporating unsupervised training data into a hierarchical system. The basic techniques we employ are the use of multiple translation models and a distinction of the hierarchical and the non-hierarchical (i.e. lexical) part of the translation model. We report experimental results on the large-scale NIST Arabic-English translation task and show that lightly-supervised training yields significant gains over the baseline.

2 Related Work

Large-scale lightly-supervised training for SMT as we define it in this paper has been first carried out by Schwenk (2008). Schwenk translates a large amount of monolingual French data with an initial Moses (Koehn et al., 2007) baseline system into English. In Schwenk's original work, an additional bilingual dictionary is added to the baseline. With lightly-supervised training, Schwenk achieves improvements of around one BLEU point over the baseline. In a later work (Schwenk and Senellart, 2009) he applies the same method for translation model adaptation on an Arabic-French task with

gains of up to 3.5 points BLEU. [1]

Hierarchical phrase-based translation has been pioneered by David Chiang (Chiang, 2005; Chiang, 2007) with his Hiero system. The hierarchical paradigm has been implemented and extended by several groups since, some have published their software as open source (Li et al., 2009; Hoang et al., 2009; Vilar et al., 2010).

Combining multiple translation models has been investigated for domain adaptation by Foster and Kuhn (2007) and Koehn and Schroeder (2007) before. Heger et al. (2010) exploit the distinction between hierarchical and lexical phrases in a similar way as we do. They train phrase translation probabilities with forced alignment using a conventional phrase-based system (Wuebker et al., 2010) and employ them for the lexical phrases while the hierarchical phrases stay untouched.

3 Using the Unsupervised Data

The most straightforward way of trying to improve the baseline with lightly-supervised training would be to concatenate the human-generated parallel data and the unsupervised data and to jointly extract phrases from the unified parallel data (after having trained word alignments for the unsupervised bitexts as well). This method is simple and expected to be effective usually. There may however be two drawbacks: First, the reliability and the amount of parallel sentences may differ between the human-generated and the unsupervised part of the training data. It might be desirable to run separate extractions on the two corpora in order to be able to distinguish and weight phrases (or rather their scores) according to their origin during decoding. Second, if we incorporate large amounts of additional unsupervised data, the amount of phrases that are extracted may become much larger. We would want to avoid blowing up our phrase table sizes without an appropriate effect on translation quality. This holds in particular in the case of hierarchical phrases. Phrase-based machine translation systems are usually able to correctly handle local context dependencies, but often have problems in producing a fluent sentence structure across long distances. It is thus an intuitive supposition that using hierarchical phrases extracted from unsupervised data in addition to the hierarchical phrases extracted from the presumably more reliable human-generated bitexts does not increase translation quality. We will compare a joint extraction to the usage of two separate translation models (either without separate weighting, with a binary feature, or as a log-linear mixture). We will further check if including hierarchical phrases from the unsupervised data is beneficial or not.

[1] Schwenk names the method *lightly-supervised training* because the topics that are covered in the monolingual source language data that is being translated may potentially also be covered by parts of the language model training data of the system which is used to translate them. This can be considered as a form of light supervision. We loosely apply the term *lightly-supervised training* if we mean the process of utilizing a machine translation system to produce additional bitexts that are used as training data, but still refer to the automatically produced bilingual corpora as *unsupervised data*.

	Arabic	English
Sentences	2 514 413	
Running words	54 324 372	55 348 390
Vocabulary	264 528	207 780
Singletons	115 171	91 390

Table 1: Data statistics for the preprocessed Arabic-English parallel training corpus. In the corpus, numerical quantities have been replaced by a special category symbol.

	dev (MT06)	test (MT08)
Sentences	1 797	1 360
Running words	49 677	45 095
Vocabulary	9 274	9 387
OOV [%]	0.5	0.4

Table 2: Data statistics for the preprocessed Arabic part of the dev and test corpora. In the corpus, numerical quantities have been replaced by a special category symbol.

4 Experiments

We use the open source Jane toolkit (Vilar et al., 2010) for our experiments, a hierarchical phrase-based translation software written in C++.

4.1 Baseline System

The baseline system has been trained using a human-generated parallel corpus of 2.5M Arabic-English sentence pairs. Word alignments in both

directions were produced with GIZA++ and symmetrized according to the refined method that was suggested by Och and Ney (2003).

The models integrated into our baseline system are: phrase translation probabilities and lexical translation probabilities for both translation directions, length penalties on word and phrase level, three binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four simple additional count- and length-based binary features, and a large 4-gram language model with modified Kneser-Ney smoothing that was trained with the SRILM toolkit (Stolcke, 2002).

We ran the cube pruning algorithm, the depth of the hierarchical recursion was restricted to one by using shallow rules as proposed by Iglesias et al. (2009).

The scaling factors of the log-linear model combination have been optimized towards BLEU with MERT (Och, 2003) on the MT06 NIST test corpus. MT08 was employed as held-out test data. Detailed statistics for the parallel training data are given in Table 1, for the development and the test corpus in Table 2.

4.2 Unsupervised Data

The unsupervised data that we integrate has been created by automatic translations of parts of the Arabic LDC Gigaword corpus (mostly from the HYT collection) with a standard phrase-based system (Koehn et al., 2003). We thus in fact conduct a cross-system and cross-paradigm variant of lightly-supervised training. Translating the monolingual Arabic data has been performed by LIUM, Le Mans, France. We thank Holger Schwenk for kindly providing the translations.

The score computed by the decoder for each translation has been normalized with respect to the sentence length and used to select the most reliable sentence pairs. Word alignments for the unsupervised data have been produced in the same way as for the baseline bilingual training data. We report the statistics of the unsupervised data in Table 3.

	Arabic	English
Sentences	4 743 763	
Running words	121 478 207	134 227 697
Vocabulary	306 152	237 645
Singletons	130 981	102 251

Table 3: Data statistics for the Arabic-English unsupervised training corpus after selection of the most reliable sentence pairs. In the corpus, numerical quantities have been replaced by a special category symbol.

4.3 Translation Models

We extracted three different phrase tables, one from the baseline human-generated parallel data only, one from the unsupervised data only, and one joint phrase table from the concatenation of the baseline data and the unsupervised data. We will denote the different extractions as *baseline*, *unsupervised*, and *joint*, respectively.

The conventional restrictions have been applied for phrase extraction in all conditions, i.e. a maximum length of ten words on source and target side for lexical phrases, a length limit of five (including non-terminal symbols) on source side and ten on target side for hierarchical phrases, and at most two non-terminals per rule which are not allowed to be adjacent on the source side. To limit the number of hierarchical phrases, a minimum count cutoff of one and an extraction pruning threshold of 0.1 have been applied to them. Note that we did not prune lexical phrases.

Statistics on the phrase table sizes are presented in Table 4.[2] In total the joint extraction results in almost three times as many phrases as the baseline extraction. The extraction from the unsupervised data exclusively results in more than two times as many hierarchical phrases as from the baseline data. The sum of the number of hierarchical phrases from baseline and unsupervised extraction is very close to the number of hierarchical phrases from the joint extraction. If we discard the hierarchical phrases extracted from the unsupervised data and use the lexical part of the unsupervised phrase table (27.3M phrases) as a second translation model in addition to the baseline phrase table (67.0M phrases), the overall number of phrases is increased by only 41% compared to the baseline system.

[2]The phrase tables have been filtered towards the phrases needed for the translation of a given collection of test corpora.

	number of phrases		
	lexical	hierarchical	total
extraction from baseline data	19.8M	47.2M	67.0M
extraction from unsupervised data	27.3M	115.6M	142.9M
phrases present in both tables	15.0M	40.1M	55.1M
joint extraction baseline + unsupervised	32.1M	166.5M	198.6M

Table 4: Phrase table sizes. The phrase tables have been filtered towards a larger set of test corpora containing a total of 2.3 million running words.

	dev (MT06)		test (MT08)	
	BLEU [%]	TER [%]	BLEU [%]	TER [%]
HPBT baseline	44.1	49.9	44.4$_{\pm 0.9}$	49.4$_{\pm 0.8}$
HPBT unsupervised only	45.3	48.8	45.2	49.1
joint extraction baseline + unsupervised	45.6	48.7	**45.4**$_{\pm 0.9}$	49.1$_{\pm 0.8}$
baseline hierarchical phrases + unsupervised lexical phrases	45.1	49.1	45.2	49.2
baseline hierarchical phrases + joint extraction lexical phrases	45.3	48.7	45.3	49.1
baseline + unsupervised lexical phrases	45.3	48.9	45.3	49.0
baseline + unsupervised lexical phrases (with binary feature)	45.3	48.8	**45.4**	49.0
baseline + unsupervised lexical phrases (separate scaling factors)	45.3	48.9	45.0	49.3
baseline + unsupervised full table	45.6	48.6	45.1	48.9
baseline + unsupervised full table (with binary feature)	45.5	48.6	45.2	48.8
baseline + unsupervised full table (separate scaling factors)	45.5	48.7	45.3	49.0

Table 5: Results for the NIST Arabic-English translation task (truecase). The 90% confidence interval is given for the baseline system as well as for the system with joint phrase extraction. Results in bold are significantly better than the baseline.

4.4 Experimental Results

The empirical evaluation of all our systems on the two standard metrics BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) is presented in Table 5. We have also checked the results for statistical significance over the baseline. The confidence intervals have been computed using bootstrapping for BLEU and Cochran's approximate ratio variance for TER (Leusch and Ney, 2009).

When we combine the full baseline phrase table with the unsupervised phrase table or the lexical part of it, we either use common scaling factors for their source-to-target and target-to-source translation costs, or we use common scaling factors but mark entries from the unsupervised table with a binary feature, or we optimize the four translation features separately for each of the two tables as part of the log-linear model combination.

Including the unsupervised data leads to a substantial gain on the unseen test set of up to +1.0% BLEU absolute. The different ways of combining the manually produced data with the unsupervised have little impact on translation quality. This holds specifically for the combination with only the lexical phrases, which, when marked with a binary feature, is able to obtain the same results as the full (joint extraction) system but with much less phrases. We compared the decoding speed of these two setups and observed that the system with less phrases is clearly faster (5.5 vs. 2.6 words per second, measured on MT08). The memory requirements of the systems do not differ greatly as we are using a binarized representation of the phrase table with on-demand loading. All setups consume slightly less than 16 gigabytes of RAM.

5 Conclusion

We presented several approaches of applying lightly-supervised training to hierarchical phrase-based machine translation. Using the additional automatically produced bitexts we have been able to obtain considerable gains compared to the baseline on the large-scale NIST Arabic-to-English translation task. We showed that a joint phrase extraction from human-generated and automatically generated parallel training data is not required to achieve significant improvements. The same translation quality can be achieved by adding a second translation model with only lexical phrases extracted from the automatically created bitexts. The overall amount of phrases can thus be kept much smaller.

Acknowledgments

The authors would like to thank Holger Schwenk from LIUM, Le Mans, France, for making the automatic translations of the Arabic LDC Gigaword corpus available. This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-08-C-0110. 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 DARPA.

References

Jean-Cédric Chappelier and Martin Rajman. 1998. A Generalized CYK Algorithm for Parsing Stochastic CFG. In *Proc. of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137, April.

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

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic, June.

Carmen Heger, Joern Wuebker, David Vilar, and Hermann Ney. 2010. A Combination of Hierarchical Systems with Forced Alignments from Phrase-Based Systems. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Paris, France, December.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 152–159, Tokyo, Japan.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 144–151, Prague, Czech Republic, June.

Gonzalo Iglesias, Adrià de Gispert, Eduardo R. Banga, and William Byrne. 2009. Rule Filtering by Pattern for Efficient Hierarchical Translation. In *Proc. of the 12th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 380–388, Athens, Greece, March.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic, June.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 177–180, Prague, Czech Republic, June.

Gregor Leusch and Hermann Ney. 2009. Edit distances with block movements and error rate confidence estimates. *Machine Translation*, December.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An Open Source Toolkit for Parsing-Based Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation*, pages 135–139, Athens, Greece, March.

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, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, pages 20–28, University of Maryland, College Park, MD, June.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

Holger Schwenk and Jean Senellart. 2009. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In *MT Summit XII*, Ottawa, Ontario, Canada, August.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 182–189, Waikiki, Hawaii, October.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, MA, August.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, volume 3, 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.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 475–484, Uppsala, Sweden, July.