Insertion and Deletion Models for Statistical Machine Translation

Matthias Huck and **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{huck,ney}@cs.rwth-aachen.de

Abstract

We investigate insertion and deletion models for hierarchical phrase-based statistical machine translation. Insertion and deletion models are designed as a means to avoid the omission of content words in the hypotheses. In our case, they are implemented as phrase-level feature functions which count the number of inserted or deleted words. An English word is considered inserted or deleted based on lexical probabilities with the words on the foreign language side of the phrase. Related techniques have been employed before by Och et al. (2003) in an n-best reranking framework and by Mauser et al. (2006) and Zens (2008) in a standard phrase-based translation system. We propose novel thresholding methods in this work and study insertion and deletion features which are based on two different types of lexicon models. We give an extensive experimental evaluation of all these variants on the NIST Chinese→English translation task.

1 Insertion and Deletion Models

In hierarchical phrase-based translation (Chiang, 2005), we deal with rules $X \rightarrow \langle \alpha, \beta, ^{\sim} \rangle$ where $\langle \alpha, \beta \rangle$ is a bilingual phrase pair that may contain symbols from a non-terminal set, i.e. $\alpha \in (\mathcal{N} \cup V_F)^+$ and $\beta \in (\mathcal{N} \cup V_E)^+$, where V_F and V_E are the source and target vocabulary, respectively, and $\mathcal{N}$ is a non-terminal set which is shared by source and target. The left-hand side of the rule is a non-terminal symbol $X \in \mathcal{N}$, and the $^{\sim}$ relation denotes a one-to-one correspondence between the non-terminals in α and in β. Let J_α denote the number of terminal symbols in α and I_β the number of terminal symbols in β. Indexing α with j, i.e. the symbol α_j, $1 \leq j \leq J_\alpha$, denotes the j-th terminal symbol on the source side of the phrase pair $\langle \alpha, \beta \rangle$, and analogous with β_i, $1 \leq i \leq I_\beta$, on the target side.

With these notational conventions, we now define our insertion and deletion models, each in both source-to-target and target-to-source direction. We give phrase-level scoring functions for the four features. In our implementation, the feature values are precomputed and written to the phrase table. The features are then incorporated directly into the log-linear model combination of the decoder.

Our insertion model in source-to-target direction $t_{\text{s2tIns}}(\cdot)$ counts the number of inserted words on the target side β of a hierarchical rule with respect to the source side α of the rule:

$$t_{\text{s2tIns}}(\alpha, \beta) = \sum_{i=1}^{I_\beta} \prod_{j=1}^{J_\alpha} \left[p(\beta_i | \alpha_j) < \tau_{\alpha_j} \right] \quad (1)$$

Here, $[\cdot]$ denotes a true or false statement: The result is 1 if the condition is true and 0 if the condition is false. The model considers an occurrence of a target word e an insertion iff no source word f exists within the phrase where the lexical translation probability $p(e|f)$ is greater than a corresponding threshold τ_f. We employ lexical translation probabilities from two different types of lexicon models, a model which is extracted from word-aligned training data and—given the word alignment matrix—relies on pure relative frequencies, and the IBM model 1 lexicon (cf. Section 2). For τ_f, previous authors have used a fixed heuristic value which was equal for all

2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 347–351,
Montréal, Canada, June 3-8, 2012.

$f \in V_f$. In Section 3, we describe how such a global threshold can be computed and set in a reasonable way based on the characteristics of the model. We also propose several novel thresholding techniques with distinct thresholds τ_f for each source word f.

In an analogous manner to the source-to-target direction, the insertion model in target-to-source direction $t_{\text{t2sIns}}(\cdot)$ counts the number of inserted words on the source side α of a hierarchical rule with respect to the target side β of the rule:

$$t_{\text{t2sIns}}(\alpha, \beta) = \sum_{j=1}^{J_\alpha} \prod_{i=1}^{I_\beta} [p(\alpha_j|\beta_i) < \tau_{\beta_i}] \quad (2)$$

Target-to-source lexical translation probabilities $p(f|e)$ are thresholded with values τ_e which may be distinct for each target word e. The model considers an occurrence of a source word f an insertion iff no target word e exists within the phrase with $p(f|e)$ greater than or equal to τ_e.

Our deletion model, compared to the insertion model, interchanges the connection of the direction of the lexical probabilities and the order of source and target in the sum and product of the term. The source-to-target deletion model thus differs from the target-to-source insertion model in that it employs a source-to-target word-based lexicon model.

The deletion model in source-to-target direction $t_{\text{s2tDel}}(\cdot)$ counts the number of deleted words on the source side α of a hierarchical rule with respect to the target side β of the rule:

$$t_{\text{s2tDel}}(\alpha, \beta) = \sum_{j=1}^{J_\alpha} \prod_{i=1}^{I_\beta} \left[p(\beta_i|\alpha_j) < \tau_{\alpha_j}\right] \quad (3)$$

It considers an occurrence of a source word f a deletion iff no target word e exists within the phrase with $p(e|f)$ greater than or equal to τ_f.

The target-to-source deletion model $t_{\text{t2sDel}}(\cdot)$ correspondingly considers an occurrence of a target word e a deletion iff no source word f exists within the phrase with $p(f|e)$ greater than or equal to τ_e:

$$t_{\text{t2sDel}}(\alpha, \beta) = \sum_{i=1}^{I_\beta} \prod_{j=1}^{J_\alpha} [p(\alpha_j|\beta_i) < \tau_{\beta_i}] \quad (4)$$

2 Lexicon Models

We restrict ourselves to the description of the source-to-target direction of the models.

2.1 Word Lexicon from Word-Aligned Data

Given a word-aligned parallel training corpus, we are able to estimate single-word based translation probabilities $p_{\text{RF}}(e|f)$ by relative frequency (Koehn et al., 2003). With $N(e, f)$ denoting counts of aligned cooccurrences of target word e and source word f, we can compute

$$p_{\text{RF}}(e|f) = \frac{N(e, f)}{\sum_{e'} N(e', f)} \; . \quad (5)$$

If an occurrence of e has multiple aligned source words, each of the alignment links contributes with a fractional count.

We denote this model as relative frequency (RF) word lexicon.

2.2 IBM Model 1

The IBM model 1 lexicon (IBM-1) is the first and most basic one in a sequence of probabilistic generative models (Brown et al., 1993). For IBM-1, several simplifying assumptions are made, so that the probability of a target sentence e_1^I given a source sentence f_0^J (with $f_0 = \text{NULL}$) can be modeled as

$$Pr(e_1^I|f_1^J) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} \sum_{j=0}^{J} p_{\text{ibm1}}(e_i|f_j) \; . \quad (6)$$

The parameters of IBM-1 are estimated iteratively by means of the Expectation-Maximization algorithm with maximum likelihood as training criterion.

3 Thresholding Methods

We introduce thresholding methods for insertion and deletion models which set thresholds based on the characteristics of the lexicon model that is applied. For all the following thresholding methods, we disregard entries in the lexicon model with probabilities that are below a fixed floor value of 10^{-6}. Again, we restrict ourselves to the description of the source-to-target direction.

individual τ_f is a distinct value for each f, computed as the arithmetic average of all entries $p(e|f)$ of any e with the given f in the lexicon model.

NIST Chinese→English	MT06 (Dev) BLEU [%]	MT06 (Dev) TER [%]	MT08 (Test) BLEU [%]	MT08 (Test) TER [%]
Baseline (with s2t+t2s RF word lexicons)	32.6	61.2	25.2	66.6
+ s2t+t2s insertion model (RF, individual)	32.9	61.4	25.7	66.2
+ s2t+t2s insertion model (RF, global)	32.8	61.8	25.7	66.7
+ s2t+t2s insertion model (RF, histogram 10)	32.9	61.7	25.5	66.5
+ s2t+t2s insertion model (RF, all)	32.8	62.0	**26.1**	66.7
+ s2t+t2s insertion model (RF, median)	32.9	62.1	25.7	67.1
+ s2t+t2s deletion model (RF, individual)	32.7	61.4	25.6	66.5
+ s2t+t2s deletion model (RF, global)	33.0	61.3	25.8	66.1
+ s2t+t2s deletion model (RF, histogram 10)	32.9	61.4	**26.0**	66.1
+ s2t+t2s deletion model (RF, all)	33.0	61.4	25.9	66.4
+ s2t+t2s deletion model (RF, median)	32.9	61.5	25.8	66.7
+ s2t+t2s insertion model (IBM-1, individual)	33.0	61.4	**26.1**	66.4
+ s2t+t2s insertion model (IBM-1, global)	33.0	61.6	25.9	66.5
+ s2t+t2s insertion model (IBM-1, histogram 10)	**33.7**	61.3	**26.2**	66.5
+ s2t+t2s insertion model (IBM-1, median)	33.0	61.3	**26.0**	66.4
+ s2t+t2s deletion model (IBM-1, individual)	32.8	61.5	**26.0**	66.2
+ s2t+t2s deletion model (IBM-1, global)	32.9	61.3	25.9	66.1
+ s2t+t2s deletion model (IBM-1, histogram 10)	32.8	61.2	25.7	66.0
+ s2t+t2s deletion model (IBM-1, median)	32.8	61.6	25.6	66.7
+ s2t insertion + s2t deletion model (IBM-1, individual)	32.7	62.3	25.7	67.1
+ s2t insertion + t2s deletion model (IBM-1, individual)	32.7	62.2	25.9	66.8
+ t2s insertion + s2t deletion model (IBM-1, individual)	33.1	61.3	25.9	66.2
+ t2s insertion + t2s deletion model (IBM-1, individual)	33.0	61.3	**26.1**	66.0
+ source+target unaligned word count	32.3	61.8	25.6	66.7
+ phrase-level s2t+t2s IBM-1 word lexicons	**33.8**	**60.5**	**26.9**	**65.4**
+ source+target unaligned word count	**34.0**	**60.4**	**26.7**	**65.8**
+ s2t+t2s insertion model (IBM-1, histogram 10)	**34.0**	**60.3**	**26.8**	**65.2**
+ phrase-level s2t+t2s DWL + triplets + discrim. RO	**34.8**	**59.8**	**27.7**	**64.7**
+ s2t+t2s insertion model (RF, individual)	**35.0**	**59.5**	**27.8**	**64.4**

Table 1: Experimental results for the NIST Chinese→English translation task (truecase). *s2t* denotes source-to-target scoring, *t2s* target-to-source scoring. Bold font indicates results that are significantly better than the baseline ($p < .1$).

global The same value $\tau_f = \tau$ is used for all f. We compute this global threshold by averaging over the individual thresholds.[1]

histogram n τ_f is a distinct value for each f. τ_f is set to the value of the $n+1$-th largest probability $p(e|f)$ of any e with the given f.

all All entries with probabilities larger than the floor value are not thresholded. This variant may be considered as *histogram* ∞. We only apply it with RF lexicons.

median τ_f is a median-based distinct value for each f, i.e. it is set to the value that separates the higher half of the entries from the lower half of the entries $p(e|f)$ for the given f.

[1]Concrete values from our experiments are: 0.395847 for the source-to-target RF lexicon, 0.48127 for the target-to-source RF lexicon. 0.0512856 for the source-to-target IBM-1, and 0.0453709 for the target-to-source IBM-1. Mauser et al. (2006) mention that they chose their heuristic thresholds for use with IBM-1 between 10^{-1} and 10^{-4}.

4 Experimental Evaluation

We present empirical results obtained with the different insertion and deletion model variants on the

Chinese→English 2008 NIST task.[2]

4.1 Experimental Setup

To set up our systems, we employ the open source statistical machine translation toolkit Jane (Vilar et al., 2010; Vilar et al., 2012), which is freely available for non-commercial use. Jane provides efficient C++ implementations for hierarchical phrase extraction, optimization of log-linear feature weights, and parsing-based decoding algorithms. In our experiments, we use the cube pruning algorithm (Huang and Chiang, 2007) to carry out the search.

We work with a parallel training corpus of 3.0M Chinese-English sentence pairs (77.5M Chinese / 81.0M English running words). The counts for the RF lexicon models are computed from a symmetrized word alignment (Och and Ney, 2003), the IBM-1 models are produced with GIZA++. When extracting phrases, we apply several restrictions, in particular a maximum length of 10 on source and target side for lexical phrases, a length limit of five (including non-terminal symbols) for hierarchical phrases, and no more than two gaps per phrase. The models integrated into the baseline are: phrase translation probabilities and RF lexical translation probabilities on phrase level, each for both translation directions, length penalties on word and phrase level, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, source-to-target and target-to-source phrase length ratios, four binary features marking phrases that have been seen more than one, two, three or five times, respectively, and an n-gram language model. The language model is a 4-gram with modified Kneser-Ney smoothing which was trained with the SRILM toolkit (Stolcke, 2002) on a large collection of English data including the target side of the parallel corpus and the LDC Gigaword v3.

Model weights are optimized against BLEU (Papineni et al., 2002) with standard Minimum Error Rate Training (Och, 2003), performance is measured with BLEU and TER (Snover et al., 2006). We employ MT06 as development set, MT08 is used as unseen test set. The empirical evaluation of all our setups is presented in Table 1.

[2] http://www.itl.nist.gov/iad/mig/tests/mt/2008/

4.2 Experimental Results

With the best model variant, we obtain a significant improvement (90% confidence) of +1.0 points BLEU over the baseline on MT08. A consistent trend towards one of the variants cannot be observed. The results on the test set with RF lexicons or IBM-1, insertion or deletion models, and (in most of the cases) with all of the thresholding methods are roughly at the same level. For comparison we also give a result with an unaligned word count model (+0.4 BLEU).

Huck et al. (2011) recently reported substantial improvements over typical hierarchical baseline setups by just including phrase-level IBM-1 scores. When we add the IBM-1 models directly, our baseline is outperformed by +1.7 BLEU. We tried to get improvements with insertion and deletion models over this setup again, but the positive effect was largely diminished. In one of our strongest setups, which includes discriminative word lexicon models (DWL), triplet lexicon models and a discriminative reordering model (discrim. RO) (Huck et al., 2012), insertion models still yield a minimal gain, though.

5 Conclusion

Our results with insertion and deletion models for Chinese→English hierarchical machine translation are twofold. On the one hand, we achieved significant improvements over a standard hierarchical baseline. We were also able to report a slight gain by adding the models to a very strong setup with discriminative word lexicons, triplet lexicon models and a discriminative reordering model. On the other hand, the positive impact of the models was mainly noticeable when we exclusively applied lexical smoothing with word lexicons which are simply extracted from word-aligned training data, which is however the standard technique in most state-of-the-art systems. If we included phrase-level lexical scores with IBM model 1 as well, the systems barely benefited from our insertion and deletion models. Compared to an unaligned word count model, insertion and deletion models perform well.

Acknowledgments

This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June.

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, USA, June.

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.

Matthias Huck, Saab Mansour, Simon Wiesler, and Hermann Ney. 2011. Lexicon Models for Hierarchical Phrase-Based Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 191–198, San Francisco, CA, USA, December.

Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. 2012. Discriminative Reordering Extensions for Hierarchical Phrase-Based Machine Translation. In *Proc. of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May.

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.

Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 103–110, Kyoto, Japan, November.

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, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2003. Syntax for Statistical Machine Translation. Technical report, Johns Hopkins University 2003 Summer Workshop on Language Engineering, Center for Language and Speech Processing, Baltimore, MD, USA, August.

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, USA, July.

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, USA, 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, USA, 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*, pages 1–20. http://dx.doi.org/10.1007/s10590-011-9120-y.

Richard Zens. 2008. *Phrase-based Statistical Machine Translation: Models, Search, Training*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, February.