N-Gram Posterior Probabilities for Statistical Machine Translation

Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{zens,ney}@cs.rwth-aachen.de

Abstract

Word posterior probabilities are a common approach for confidence estimation in automatic speech recognition and machine translation. We will generalize this idea and introduce n-gram posterior probabilities and show how these can be used to improve translation quality. Additionally, we will introduce a sentence length model based on posterior probabilities.

We will show significant improvements on the Chinese-English NIST task. The absolute improvements of the BLEU score is between 1.1% and 1.6%.

1 Introduction

The use of word posterior probabilities is a common approach for confidence estimation in automatic speech recognition, e.g. see (Wessel, 2002). This idea has been adopted to estimate confidences for machine translation, e.g. see (Blatz et al., 2003; Ueffing et al., 2003; Blatz et al., 2004). These confidence measures were used in the computer assisted translation (CAT) framework, e.g. (Gandrabur and Foster, 2003). The (simplified) idea is that the confidence measure is used to decide if the machine-generated prediction should be suggested to the human translator or not.

There is only few work on how to improve machine translation performance using confidence measures. The only work, we are aware of, is (Blatz et al., 2003). The outcome was that the confidence measures did not result in improvements of the translation quality measured with the BLEU and NIST scores. Here, we focus on how the ideas and methods commonly used for confidence estimation can be adapted and/or extended to improve translation quality.

So far, always word-level posterior probabilities were used. Here, we will generalize this idea to n-grams.

In addition to the n-gram posterior probabilities, we introduce a sentence-length model based on posterior probabilities. The common phrase-based translation systems, such as (Och et al., 1999; Koehn, 2004), do not use an explicit sentence length model. Only the simple word penalty goes into that direction. It can be adjusted to prefer longer or shorter translations. Here, we will explicitly model the sentence length.

The novel contributions of this work are to introduce n-gram posterior probabilities and sentence length posterior probabilities. Using these methods, we achieve significant improvements of translation quality.

The remaining part of this paper is structured as follows: first, we will briefly describe the baseline system, which is a state-of-the-art phrase-based statistical machine translation system. Then, in Section 3, we will introduce the n-gram posterior probabilities. In Section 4, we will define the sentence length model. Afterwards, in Section 5, we will describe how these novel models can be used for rescoring/reranking. The experimental results will be presented in Section 6. Future applications will be described in Section 7. Finally, we will conclude in Section 8.

Proceedings of the Workshop on Statistical Machine Translation, pages 72–77,
New York City, June 2006. ©2006 Association for Computational Linguistics

2 Baseline System

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \ldots f_j \ldots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \ldots e_i \ldots e_I$. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} \qquad (1)$$

The posterior probability $Pr(e_1^I|f_1^J)$ is modeled directly using a log-linear combination of several models (Och and Ney, 2002):

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum\limits_{I',e'^{I'}_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \qquad (2)$$

The denominator is a normalization factor that depends only on the source sentence f_1^J. Therefore, we can omit it during the search process. As a decision rule, we obtain:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\} \qquad (3)$$

This approach is a generalization of the source-channel approach (Brown et al., 1990). It has the advantage that additional models $h(\cdot)$ can be easily integrated into the overall system. The model scaling factors λ_1^M are trained with respect to the final translation quality measured by an error criterion (Och, 2003).

We use a state-of-the-art phrase-based translation system as described in (Zens and Ney, 2004; Zens et al., 2005). The baseline system includes the following models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty and a phrase penalty.

3 N-Gram Posterior Probabilities

The idea is similar to the word posterior probabilities: we sum the sentence posterior probabilities for each occurrence of an n-gram.

Let $\delta(\cdot, \cdot)$ denote the Kronecker function. Then, we define the fractional count $C(e_1^n, f_1^J)$ of an n-gram e_1^n for a source sentence f_1^J as:

$$C(e_1^n, f_1^J) = \sum_{I,e'^I_1} \sum_{i=1}^{I-n+1} p(e'^I_1|f_1^J) \cdot \delta(e'^{i+n-1}_i, e_1^n) \qquad (4)$$

The sums over the target language sentences are limited to an N-best list, i.e. the N best translation candidates according to the baseline model. In this equation, the term $\delta(e'^{i+n-1}_i, e_1^n)$ is one if and only if the n-gram e_1^n occurs in the target sentence e'^I_1 starting at position i.

Then, the posterior probability of an n-gram is obtained as:

$$p(e_1^n|f_1^J) = \frac{C(e_1^n, f_1^J)}{\sum\limits_{e'^n_1} C(e'^n_1, f_1^J)} \qquad (5)$$

Note that the widely used word posterior probability is obtained as a special case, namely if n is set to one.

4 Sentence Length Posterior Probability

The common phrase-based translation systems, such as (Och et al., 1999; Koehn, 2004), do not use an explicit sentence length model. Only the simple word penalty goes into that direction. It can be adjusted to prefer longer or shorter translations.

Here, we will use the posterior probability of a specific target sentence length I as length model:

$$p(I|f_1^J) = \sum_{e_1^I} p(e_1^I|f_1^J) \qquad (6)$$

Note that the sum is carried out only over target sentences e_1^I with the a specific length I. Again, the candidate target language sentences are limited to an N-best list.

5 Rescoring/Reranking

A straightforward application of the posterior probabilities is to use them as additional features in a rescoring/reranking approach (Och et al., 2004). The use of N-best lists in machine translation has several advantages. It alleviates the effects of the huge search space which is represented in word graphs by using a compact excerpt of the N best hypotheses generated by the system. N-best lists are suitable for easily applying several rescoring techniques since the hypotheses are already fully generated. In comparison, word graph rescoring techniques need specialized tools which can traverse the graph accordingly.

The n-gram posterior probabilities can be used similar to an n-gram language model:

$$h_n(f_1^J, e_1^I) = \frac{1}{I} \log \left(\prod_{i=1}^{I} p(e_i | e_{i-n+1}^{i-1}, f_1^J) \right) \quad (7)$$

with:

$$p(e_i | e_{i-n+1}^{i-1}, f_1^J) = \frac{C(e_{i-n+1}^{i}, f_1^J)}{C(e_{i-n+1}^{i-1}, f_1^J)} \quad (8)$$

Note that the models do not require smoothing as long as they are applied to the same N-best list they are trained on.

If the models are used for unseen sentences, smoothing is important to avoid zero probabilities. We use a linear interpolation with weights α_n and the smoothed $(n-1)$-gram model as generalized distribution.

$$\begin{aligned} p_n(e_i | e_{i-n+1}^{i-1}, f_1^J) &= \alpha_n \cdot \frac{C(e_{i-n+1}^{i}, f_1^J)}{C(e_{i-n+1}^{i-1}, f_1^J)} \\ &\quad + (1-\alpha_n) \cdot p_{n-1}(e_i | e_{i-n+2}^{i-1}, f_1^J) \end{aligned} \quad (9)$$

Note that absolute discounting techniques that are often used in language modeling cannot be applied in a straightforward way, because here we have *fractional* counts.

The usage of the sentence length posterior probability for rescoring is even simpler. The resulting feature is:

$$h_L(f_1^J, e_1^I) = \log p(I | f_1^J) \quad (10)$$

Again, the model does not require smoothing as long as it is applied to the same N-best list it is trained on. If it is applied to other sentences, smoothing becomes important. We propose to smooth the sentence length model with a Poisson distribution.

$$p_\beta(I | f_1^J) = \beta \cdot p(I | f_1^J) + (1-\beta) \cdot \frac{\lambda^I \exp(-\lambda)}{I!} \quad (11)$$

We use a linear interpolation with weight β. The mean λ of the Poisson distribution is chosen to be identical to the mean of the unsmoothed length model:

$$\lambda = \sum_I I \cdot p(I | f_1^J) \quad (12)$$

6 Experimental Results

6.1 Corpus Statistics

The experiments were carried out on the large data track of the Chinese-English NIST task. The corpus statistics of the bilingual training corpus are shown in Table 1. The language model was trained on the English part of the bilingual training corpus and additional monolingual English data from the GigaWord corpus. The total amount of language model training data was about 600M running words. We use a fourgram language model with modified Kneser-Ney smoothing as implemented in the SRILM toolkit (Stolcke, 2002).

To measure the translation quality, we use the BLEU score (Papineni et al., 2002) and the NIST score (Doddington, 2002). The BLEU score is the geometric mean of the n-gram precision in combination with a brevity penalty for too short sentences. The NIST score is the arithmetic mean of a weighted n-gram precision in combination with a brevity penalty for too short sentences. Both scores are computed case-sensitive with respect to four reference translations using the mteval-v11b tool[1]. As the BLEU and NIST scores measure accuracy higher scores are better.

We use the BLEU score as primary criterion which is optimized on the development set using the Downhill Simplex algorithm (Press et al., 2002). As development set, we use the NIST 2002 evaluation set. Note that the baseline system is already well-tuned and would have obtained a high rank in the last NIST evaluation (NIST, 2005).

6.2 Translation Results

The translation results for the Chinese-English NIST task are presented in Table 2. We carried out experiments for evaluation sets of several years. For these rescoring experiments, we use the 10 000 best translation candidates, i.e. N-best lists of size N=10 000.

[1] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

Table 1: Chinese-English NIST task: corpus statistics for the bilingual training data and the NIST evaluation sets of the years 2002 to 2005.

			Chinese	English
Train	Sentence Pairs		7M	
	Running Words		199M	213M
	Vocabulary Size		223K	351K
	Dictionary Entry Pairs		82K	
Eval	2002	Sentences	878	3 512
		Running Words	25K	105K
	2003	Sentences	919	3 676
		Running Words	26K	122K
	2004	Sentences	1788	7 152
		Running Words	52K	245K
	2005	Sentences	1082	4 328
		Running Words	33K	148K

Using the 1-gram posterior probabilities, i.e. the conventional word posterior probabilities, there is only a very small improvement, or no improvement at all. This is consistent with the findings of the JHU workshop on confidence estimation for statistical machine translation 2003 (Blatz et al., 2003), where the word-level confidence measures also did not help to improve the BLEU or NIST scores.

Successively adding higher order n-gram posterior probabilities, the translation quality improves consistently across all evaluation sets. We also performed experiments with n-gram orders beyond four, but these did not result in further improvements.

Adding the sentence length posterior probability feature is also helpful for all evaluation sets. For the development set, the overall improvement is 1.5% for the BLEU score. On the blind evaluation sets, the overall improvement of the translation quality ranges between 1.1% and 1.6% BLEU.

Some translation examples are shown in Table 3.

7 Future Applications

We have shown that the n-gram posterior probabilities are very useful in a rescoring/reranking framework. In addition, there are several other potential applications. In this section, we will describe two of them.

7.1 Iterative Search

The n-gram posterior probability can be used for rescoring as described in Section 5. An alternative is to use them directly during the search. In this second search pass, we use the models from the first pass, i.e. the baseline system, and additionally the n-gram and sentence length posterior probabilities. As the n-gram posterior probabilities are basically a kind of sentence-specific language model, it is straightforward to integrate them. This process can also be iterated. Thus, using the N-best list of the second pass to recompute the n-gram and sentence length posterior probabilities and do a third search pass, etc..

7.2 Computer Assisted Translation

In the computer assisted translation (CAT) framework, the goal is to improve the productivity of human translators. The machine translation system takes not only the current source language sentence but also the already typed partial translation into account. Based on this information, the system suggest completions of the sentence. Word-level posterior probabilities have been used to select the most appropriate completion of the system, for more details see e.g. (Gandrabur and Foster, 2003; Ueffing and Ney, 2005). The n-gram based posterior probabilities as described in this work, might be better suited for this task as they explicitly model the dependency on the previous words, i.e. the given prefix.

8 Conclusions

We introduced n-gram and sentence length posterior probabilities and demonstrated their usefulness for rescoring purposes. We performed systematic experiments on the Chinese-English NIST task and showed significant improvements of the translation quality. The improvements were consistent among several evaluation sets.

An interesting property of the introduced methods is that they do not require additional knowledge sources. Thus the given knowledge sources are better exploited. Our intuition is that the posterior models prefer hypotheses with n-grams that are common in the N-best list.

The achieved results are promising. Despite that, there are several ways to improve the approach.

Table 2: Case-sensitive translation results for several evaluation sets of the Chinese-English NIST task.

Evaluation set	2002 (dev)		2003		2004		2005	
System	NIST	BLEU[%]	NIST	BLEU[%]	NIST	BLEU[%]	NIST	BLEU[%]
Baseline	8.49	30.5	8.04	29.5	8.14	29.0	8.01	28.2
+ 1-grams	8.51	30.5	8.08	29.5	8.17	29.0	8.03	28.2
+ 2-grams	8.47	30.8	8.03	29.7	8.12	29.2	7.98	28.1
+ 3-grams	8.73	31.6	8.25	30.1	8.45	30.0	8.20	28.6
+ 4-grams	8.74	31.7	8.26	30.1	8.47	30.1	8.20	28.6
+ length	8.87	32.0	8.42	30.9	8.60	30.6	8.34	29.3

Table 3: Translation examples for the Chinese-English NIST task.

Baseline	At present, there is no organization claimed the attack.
Rescored	At present, there is no organization claimed responsibility for the attack.
Reference	So far, no organization whatsoever has claimed responsibility for the attack.
Baseline	FIFA to severely punish football fraud
Rescored	The International Football Federation (FIFA) will severely punish football's deception
Reference	FIFA will severely punish all cheating acts in the football field
Baseline	In more than three months of unrest, a total of more than 60 dead and 2000 injured.
Rescored	In more than three months of unrest, a total of more than 60 people were killed and more than 2000 injured.
Reference	During the unrest that lasted more than three months, a total of more than 60 people died and over 2,000 were wounded.

For the decision rule in Equation 3, the model scaling factors λ_1^M can be multiplied with a constant factor without changing the result. This global factor would affect the proposed posterior probabilities. So far, we have not tuned this parameter, but a proper adjustment might result in further improvements.

Currently, the posterior probabilities are computed on an N-best list. Using word graphs instead should result in more reliable estimates, as the number of hypotheses in a word graph is some orders of a magnitude larger than in an N-best list.

Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org).

References

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence estimation for machine translation. Final report, JHU/CLSP Summer Workshop. `http://www.clsp.jhu.edu/ws2003/groups/estimate/`.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence estimation for machine translation. In *Proc. 20th Int. Conf. on Computational Linguistics (COLING)*, pages 315–321, Geneva, Switzerland, August.

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

G. Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. ARPA Workshop on Human Language Technology*.

S. Gandrabur and G. Foster. 2003. Confidence estimation for text prediction. In *Proc. Conf. on Natural Lan-*

guage Learning (CoNLL), pages 95–102, Edmonton, Canada, May.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *6th Conf. of the Association for Machine Translation in the Americas (AMTA 04)*, pages 115–124, Washington DC, September/October.

NIST. 2005. NIST 2005 machine translation evaluation official results. `http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html`, August.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

F. J. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

F. J. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 161–168, Boston,MA.

F. J. 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.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

N. Ueffing and H. Ney. 2005. Application of word-level confidence measures in interactive statistical machine translation. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, pages 262–270, Budapest, Hungary, May.

N. Ueffing, K. Macherey, and H. Ney. 2003. Confidence Measures for Statistical Machine Translation. In *Proc. MT Summit IX*, pages 394–401, New Orleans, LA, September.

F. Wessel. 2002. *Word Posterior Probabilities for Large Vocabulary Continuous Speech Recognition*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, January.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. Human Language Technology Conf. / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 257–264, Boston, MA, May.

R. Zens, O. Bender, S. Hasan, S. Khadivi, E. Matusov, J. Xu, Y. Zhang, and H. Ney. 2005. The RWTH phrase-based statistical machine translation system. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 155–162, Pittsburgh, PA, October.