Exploiting N-best Hypotheses for SMT Self-Enhancement

Boxing Chen, Min Zhang, Aiti Aw and **Haizhou Li**
Department of Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace, 119613, Singapore
{bxchen, mzhang, aaiti, hli}@i2r.a-star.edu.sg

Abstract

Word and n-gram posterior probabilities estimated on N-best hypotheses have been used to improve the performance of statistical machine translation (SMT) in a rescoring framework. In this paper, we extend the idea to estimate the posterior probabilities on N-best hypotheses for translation phrase-pairs, target language n-grams, and source word reorderings. The SMT system is self-enhanced with the posterior knowledge learned from N-best hypotheses in a re-decoding framework. Experiments on NIST Chinese-to-English task show performance improvements for all the strategies. Moreover, the combination of the three strategies achieves further improvements and outperforms the baseline by 0.67 BLEU score on NIST-2003 set, and 0.64 on NIST-2005 set, respectively.

1 Introduction

State-of-the-art Statistical Machine Translation (SMT) systems usually adopt a two-pass search strategy. In the first pass, a decoding algorithm is applied to generate an N-best list of translation hypotheses; while in the second pass, the final translation is selected by rescoring and re-ranking the N-best hypotheses through additional feature functions. In this framework, the N-best hypotheses serve as the candidates for the final translation selection in the second pass.

These N-best hypotheses can also provide useful feedback to the MT system as the first decoding has discarded many undesirable translation candidates. Thus, the knowledge captured in the N-best hypotheses, such as *posterior probabilities* for words, n-grams, phrase-pairs, and source word reorderings, etc. is more compatible with the source sentences and thus could potentially be used to improve the translation performance.

Word posterior probabilities estimated from the N-best hypotheses have been widely used for confidence measure in automatic speech recognition (Wessel, 2002) and have also been adopted into machine translation. Blatz et al. (2003) and Ueffing et al. (2003) used word posterior probabilities to estimate the confidence of machine translation. Chen et al. (2005), Zens and Ney (2006) reported performance improvements by computing target n-grams posterior probabilities estimated on the N-best hypotheses in a rescoring framework. Transductive learning method (Ueffing et al., 2007) which repeatedly re-trains the generated source-target N-best hypotheses with the original training data again showed translation performance improvement and demonstrated that the translation model can be reinforced from N-best hypotheses.

In this paper, we further exploit the potential of the N-best hypotheses and propose several schemes to derive the posterior knowledge from the N-best hypotheses, in an effort to enhance the language model, translation model, and source word reordering under a re-decoding framework of any phrase-based SMT system.

2 Self-Enhancement with Posterior Knowledge

The self-enhancement system structure is shown in Figure 1. Our baseline system is set up using Moses (Koehn et al., 2007), a state-of-the-art phrase-base SMT open source package. In the followings, we detail the approaches to exploiting the three different kinds of posterior knowledge, namely, language model, translation model and word reordering.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 157–160,
Columbus, Ohio, USA, June 2008.

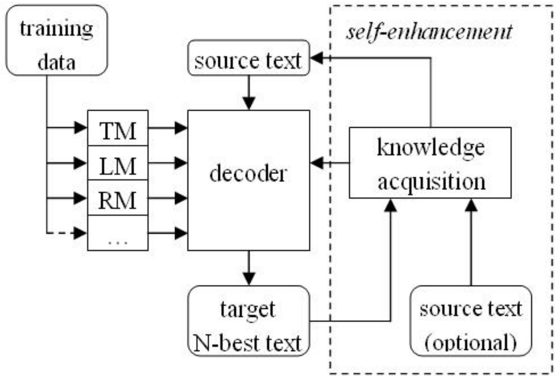

Figure 1: Self-enhancement system structure, where TM is translation model, LM is language model, and RM is reordering model.

2.1 Language Model

We consider self-enhancement of language model as a language model adaptation problem similar to (Nakajima et al., 2002). The original monolingual target training data is regarded as general-domain data while the test data as a domain-specific data. Obviously, the real domain-specific target data (test data) is unavailable for training. In this work, the N-best hypotheses of the test set are used as a quasi-corpus to train a language model. This new language model trained on the quasi-corpus is then used together with the language model trained on the general-domain data (original training data) to produce a new list of N-best hypotheses under our self-enhancement framework. The feature function of the language model $h_{LM}(f_1^J, e_1^I)$ is a mixture model of the two language models as in Equation 1.

$$h_{LM}(f_1^J, e_1^I) = \lambda_1 h_{TLM}(e_1^I) + \lambda_2 h_{QLM}(e_1^I) \qquad (1)$$

where f_1^J is the source language words string, e_1^I is the target language words string, *TLM* is the language model trained on target training data, and *QLM* is on the quasi-corpus of N-best hypotheses.

The mixture model exploits multiple language models with weights λ_1 and λ_2 being optimized together with other feature functions. The procedure for self-enhancement of the language model is as follows:

1. Run decoding and extract N-best hypotheses.
2. Train a new language model (*QLM*) on the N-best hypotheses.
3. Optimize the weights of the decoder which uses both original LM (*TLM*) and the new LM (*QLM*).
4. Repeat step 1-3 for a fixed number of iterations.

2.2 Translation Model

In general, we can safely assume that for a given source input, phrase-pairs that appeared in the N-best hypotheses are better than those that did not. We call the former "good phrase-pairs" and the later "bad phrase-pairs" for the given source input. Hypothetically, we can reinforce the translation model by appending the "good phrase-pairs" to the original phrase table and changing the probability space of the translation model, as phrase-based translation probabilities are estimated using relative frequencies. The new direct phrase-based translation probabilities are computed as follows:

$$p(\tilde{e} \mid \tilde{f}) = \frac{N_{train}(\tilde{f}, \tilde{e}) + N_{nbest}(\tilde{f}, \tilde{e})}{N_{train}(\tilde{f}) + N_{nbest}(\tilde{f})} \qquad (2)$$

where $\tilde{f}$ is the source language phrase, $\tilde{e}$ is the target language phrase, $N_{train}(.)$ is the frequencies observed in the training data, and $N_{nbest}(.)$ is the frequencies observed in the N-best hypotheses. For those phrase-pairs that did not appear in the N-best hypotheses list ("bad phrase-pairs"), $N_{nbest}(\tilde{f}, \tilde{e})$ equals 0, but the marginal count of $\tilde{f}$ is increased by $N_{nbest}(\tilde{f})$, in this way the phrase-based translation probabilities of "bad phrase-pairs" degraded when compared with the corresponding probabilities in the original translation model, and that of "good phrase-pairs" increased, hence improve the translation model.

The procedure for translation model self-enhancement can be summarized as follows:

1. Run decoding and extract N-best hypotheses.
2. Extract "good phrase-pairs" according to the hypotheses' phrase-alignment information and append them to the original phrase table to generate a new phrase table.
3. Score the new phrase table to create a new translation model.
4. Optimize the weights of the decoder with the above new translation model.
5. Repeat step 1-4 for a fixed number of iterations.

2.3 Word Reordering

Some previous work (Costa-jussà and Fonollosa, 2006; Li et al., 2007) have shown that reordering a source sentence to match the word order in its corresponding target sentence can produce better translations for a phrase-based SMT system. We bring this idea forward to our word reordering self-enhancement framework, which similarly translates a source sentence (*S*) to target sentence (*T*) in two stages: $S \rightarrow S' \rightarrow T$, where S' is the reordered source sentence.

The phrase-alignment information in each hypothesis indicates the word reordering for source sentence. We select the word reordering with the highest posterior probability as the best word reordering for a given source sentence. Word reorderings from different phrase segmentation but with same word surface order are merged. The posterior probabilities of the word re-orderings are computed as in Equation 3.

$$p(r_1^J \mid f_1^J) = \frac{N(r_1^J)}{N_{hyp}} \qquad (3)$$

where $N(r_1^J)$ is the count of word reordering r_1^J, and N_{hyp} is the number of N-best hypotheses.

The words of the source sentence are then reordered according to their indices in the best selected word reordering r_1^J. The procedure for self-enhancement of word reordering is as follows.

1. Run decoding and extract N-best hypotheses.
2. Select the best word re-orderings according to the phrase-alignment information.
3. Reorder the source sentences according to the selected word reordering.
4. Optimize the weights of the decoder with the reordered source sentences.
5. Repeat step 1-4 for a fixed number of iterations.

3 Experiments and Results

Experiments on Chinese-to-English NIST translation tasks were carried out on the FBIS[1] corpus. We used NIST 2002 MT evaluation test set as our development set, and the NIST 2003, 2005 test sets as our test sets as shown in Table 1.

We determine the number of iteration empirically by setting it to 10. We then observe the BLEU score on the development set for each iteration. The iteration number which achieved the best BLEU score on development set is selected as the iteration number of iterations for the test set.

Data set	type	#Running words	
		Chinese	English
train	parallel	7.0M	8.9M
	monolingual	-	61.5M
NIST 02	dev	23.2K	108.6K
NIST 03	test	25.8K	116.5K
NIST 05	test	30.5K	141.9K

Table 1: Statistics of training, dev and test sets. Evaluation sets of NIST campaigns include 4 references: total numbers of running words are provided in the table.

System	#iter.	NIST 02	NIST 03	NIST 05
Base	-	27.67	26.68	24.82
TM	4	27.87	26.95	25.05
LM	6	27.96	27.06	25.07
WR	6	27.99	27.04	25.11
Comb	7	**28.45**	**27.35**	**25.46**

Table 2: BLEU% scores of five systems: decoder (Base), self-enhancement on translation model (TM), language model (LM), word reordering (WR) and the combination of TM, LM and WR (Comb).

Further experiments also suggested that, in this experiment scenario, setting the size of N-best list to 3,000 arrives at the greatest performance improvements. Our evaluation metric is BLEU (Papineni et al., 2002). The translation performance is reported in Table 2, where the column "#iter." refers to the iteration number where the system achieved the best BLEU score on development set.

Compared with the baseline ("Base" in Table 2), all three self-enhancement methods ("TM", "LM", and "WR" in Table 2) consistently improved the performance. In general, absolute gains of 0.23-0.38 BLEU score were obtained for each method on two test sets. While comparing the performance among all three methods, we can see that they achieved very similar improvement. Combining the three methods showed further gains in BLEU score. Totally, the combined system outperformed the baseline by 0.67 BLEU score on NIST'03, and 0.64 on NIST'05 test set, respectively.

4 Discussion

As posterior knowledge applied in our models are *posterior probabilities*, the main difference between our work and all previous work is the use of knowledge source, where we derive knowledge from the N-best hypotheses generated from previous iteration.

[1] LDC2003E14

Comparing the work of (Nakajima et al., 2002), there is a slight difference between the two models. Nakajima et al. used only 1-best hypothesis, while we use N-best hypotheses of test set as the quasi-corpus to train the language model.

In the work of (Costa-jussà and Fonollosa, 2006; Li et al., 2007) which similarly translates a source sentence (S) to target sentence (T) in two stages: $S \rightarrow S' \rightarrow T$, they derive S' from training data; while we obtain S' based on the occurrence frequency, i.e. posterior probability of each source word reordering in the N-best hypotheses list.

An alternative solution for enhancing the translation model is through self-training (Ueffing, 2006; Ueffing et al., 2007) which re-trains the source-target N-best hypotheses together with the original training data, and thus differs from ours in the way of new phrase pairs extraction. We only supplement those phrase-pairs appeared in the N-best hypotheses to the original phrase table. Further experiment showed that improvement obtained by self-training method is not as consistent on both development and test sets as that by our method. One possible reason is that in self-training, the entire translation model is adjusted with the addition of new phrase-pairs extracted from the source-target N-best hypotheses, and hence the effect is less predictable.

5 Conclusions

To take advantage of the N-best hypotheses, we proposed schemes in a re-decoding framework and made use of the posterior knowledge learned from the N-best hypotheses to improve a phrase-based SMT system. The posterior knowledge include posterior probabilities for target n-grams, translation phrase-pairs and source word re-orderings, which in turn improve the language model, translation model, and word reordering respectively.

Experiments were based on the state-of-the-art phrase-based decoder and carried out on NIST Chinese-to-English task. It has been shown that all three methods improved the performance. Moreover, the combination of all three strategies outperforms each individual method and significantly outperforms the baseline. We demonstrated that the SMT system can be self-enhanced by exploiting useful feedback from the N-best hypotheses which are generated by itself.

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.*

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. In *Proceeding of IWSLT-2005*, pp.98-104, Pittsburgh, USA, October.

M. R. Costa-jussà, J. A. R. Fonollosa. 2006. Statistical Machine Reordering. In *Proceeding of EMNLP 2006.*

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL-2007*, pp. 177-180, Prague, Czech Republic.

C.-H. Li, M. Li, D. Zhang, M. Li, M. Zhou and Y. Guan. 2007. A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. In *Proceedings of ACL-2007.* Prague, Czech Republic.

H. Nakajima, H. Yamamoto, T. Watanabe. 2002. Language model adaptation with additional text generated by machine translation. In *Proceedings of COLING-2002*. Volume 1, Pages: 1-7. Taipei.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceeding of ACL-2002*, pp. 311-318.

N. Ueffing. 2006. Using Monolingual Source-Language Data to Improve MT Performance. In *Proceedings of IWSLT 2006*. Kyoto, Japan. November 27-28.

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

N. Ueffing, G. Haffari, A. Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of ACL-2007*, Prague.

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. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Proceedings of the HLT-NAACL Workshop on SMT*, pp. 72-77, NY.