Fast Consensus Hypothesis Regeneration for Machine Translation

Boxing Chen, George Foster and **Roland Kuhn**
National Research Council Canada
283 Alexandre-Taché Boulevard, Gatineau (Québec), Canada J8X 3X7
{Boxing.Chen, George.Foster, Roland.Kuhn}@nrc.ca

Abstract

This paper presents a fast consensus hypothesis regeneration approach for machine translation. It combines the advantages of feature-based fast consensus decoding and hypothesis regeneration. Our approach is more efficient than previous work on hypothesis regeneration, and it explores a wider search space than consensus decoding, resulting in improved performance. Experimental results show consistent improvements across language pairs, and an improvement of up to 0.72 BLEU is obtained over a competitive single-pass baseline on the Chinese-to-English NIST task.

1 Introduction

State-of-the-art statistical machine translation (SMT) systems are often described as a two-pass process. In the first pass, decoding algorithms are applied to generate either a translation *N*-best list or a translation forest. Then in the second pass, various re-ranking algorithms are adopted to compute the final translation. The re-ranking algorithms include rescoring (Och et al., 2004) and Minimum Bayes-Risk (MBR) decoding (Kumar and Byrne, 2004; Zhang and Gildea, 2008; Tromble et al., 2008). Rescoring uses more sophisticated additional feature functions to score the hypotheses. MBR decoding directly incorporates the evaluation metrics (i.e., loss function), into the decision criterion, so it is effective in tuning the MT performance for a specific loss function. In particular, sentence-level BLEU loss function gives gains on BLEU (Kumar and Byrne, 2004).

The naïve MBR algorithm computes the loss function between every pair of k hypotheses, needing $O(k^2)$ comparisons. Therefore, only small number k is applicable. Very recently, DeNero et al. (2009) proposed a fast consensus decoding (FCD) algorithm in which the similarity scores are computed based on the feature expectations over the translation *N*-best list or translation forest. It is equivalent to MBR decoding when using a linear similarity function, such as unigram precision.

Re-ranking approaches improve performance on an *N*-best list whose contents are fixed. A complementary strategy is to augment the contents of an *N*-best list in order to broaden the search space. Chen et al (2008) have proposed a three-pass SMT process, in which a hypothesis regeneration pass is added between the decoding and rescoring passes. New hypotheses are generated based on the original *N*-best hypotheses through *n*-gram expansion, confusion-network decoding or re-decoding. All three hypothesis regeneration methods obtained decent and comparable improvements in conjunction with the same rescoring model. However, since the final translation candidates in this approach are produced from different methods, local feature functions (such as translation models and reordering models) of each hypothesis are not directly comparable and rescoring must exploit rich global feature functions to compensate for the loss of local feature functions. Thus this approach is dependent on the use of computationally expensive features for rescoring, which makes it inefficient.

In this paper, we propose a fast consensus hypothesis regeneration method that combines the advantages of feature-based fast consensus decoding and hypothesis regeneration. That is, we integrate the feature-based similarity/loss function based on evaluation metrics such as BLEU score into the hypothesis regeneration procedure to score the partial hypotheses in the beam search and compute the final translations. Thus, our approach is more efficient than the original three-pass hypothesis regeneration. Moreover, our approach explores more search space than consen-

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 11–16,
Uppsala, Sweden, 15-16 July 2010.

sus decoding, giving it an advantage over the latter.

In particular, we extend linear corpus BLEU (Tromble et al., 2008) to *n*-gram expectation-based linear BLEU, then further extend the *n*-gram expectation computed on full-length hypotheses to n-gram expectation computed on fixed-length partial hypotheses. Finally, we extend the hypothesis regeneration with forward *n*-gram expansion to bidirectional *n*-gram expansion including both the forward and backward *n*-gram expansion. Experimental results show consistent improvements over the baseline across language pairs, and up to 0.72 BLEU points are obtained from a competitive baseline on the Chinese-to-English NIST task.

2 Fast Consensus Hypothesis Regeneration

Since the three hypothesis regeneration methods with *n*-gram expansion, confusion network decoding and re-decoding produce very similar performance (Chen et al., 2008), we consider only *n*-gram expansion method in this paper. *N*-gram expansion can (almost) fully exploit the search space of target strings which can be generated by an *n*-gram language model trained on the *N*-best hypotheses (Chen et al., 2007).

2.1 Hypothesis regeneration with bidirectional n-gram expansion

N-gram expansion (Chen et al., 2007) works as follows: firstly, train an *n*-gram language model based on the translation *N*-best list or translation forest; secondly, expand each partial hypothesis by appending a word via overlapped (*n*-1)-grams until the partial hypothesis reaches the sentence ending symbol. In each expanding step, the partial hypotheses are pruned through a beam-search algorithm with scoring functions.

Duchateau et al. (2001) shows that the backward language model contains information complementary to the information in the forward language model. Hence, on top of the forward *n*-gram expansion used in (Chen et al., 2008), we further introduce backward *n*-gram expansion to the hypothesis regeneration procedure. Backward *n*-gram expansion involves letting the partial hypotheses start from the last words that appeared in the translation *N*-best list and having the expansion go from right to left.

Figure 1 gives an example of backward *n*-gram expansion. The second row shows bi-grams which are extracted from the original hypotheses in the first row. The third row shows how a partial hypothesis is expanded via backward *n*-gram expansion method. The fourth row lists some new hypotheses generated by backward *n*-gram expansion which do not exist in the original hypothesis list.

original hypotheses	*about weeks' work .*			
	one week's work			
	about one week's			
	about a week work			
	about one week work			
bi-grams	*about weeks', weeks' work, …, about one, …, week work.*			
backward n-gram expansion	partial hyp.		***week's***	*work*
	n-gram	*one*	***week's***	
	new partial hyp.	*one*	***week's***	*work*
new hypotheses	***about one week's work***			
	about week's work			
	one weeks' work .			
	one week's work .			
	one week's work .			

Figure 1: Example of original hypotheses; bi-grams collected from them; backward expanding a partial hypothesis via an overlapped *n*-1-gram; and new hypotheses generated through backward *n*-gram expansion.

2.2 Feature-based scoring functions

To speed up the search, the partial hypotheses are pruned via beam-search in each expanding step. Therefore, the scoring functions applied with the beam-search algorithm are very important. In (Chen et al., 2008), more than 10 additional global features are computed to rank the partial hypothesis list, and this is not an efficient way. In this paper, we propose to directly incorporate the evaluation metrics such as BLEU score to rank the candidates. The scoring functions of this work are derived from the method of lattice Minimum Bayes-risk (MBR) decoding (Tromble et al., 2008) and fast consensus decoding (DeNero et al., 2009), which were originally inspired from *N*-best MBR decoding (Kumar and Byrne, 2004).

From a set of translation candidates *E,* MBR decoding chooses the translation that has the least expected loss with respect to other candidates. Given a hypothesis set *E*, under the probability model $P(e \mid f)$, MBR computes the translation $\tilde{e}$ as follows:

$$\tilde{e} = \arg\min_{e \in E} \sum_{e' \in E} L(e, e') \cdot P(e \mid f) \quad (1)$$

where f is the source sentence, $L(e, e')$ is the loss function of two translations e and e'.

Suppose that we are interested in maximizing the BLEU score (Papineni et al., 2002) to optimize the translation performance. The loss function is defined as $L(e, e') = 1 - BLEU(e, e')$, then the MBR objective can be re-written as

$$\tilde{e} = \arg\max_{e \in E} \sum_{e' \in E} BLEU(e, e') \cdot P(e \mid f) \quad (2)$$

E represents the space of the translations. For N-best MBR decoding, this space is the N-best list produced by a baseline decoder (Kumar and Byrne, 2004). For lattice MBR decoding, this space is the set of candidates encoded in the lattice (Tromble et al., 2008). Here, with hypothesis regeneration, this space includes: 1) the translations produced by the baseline decoder either in an N-best list or encoded in a translation lattice, and 2) the translations created by hypothesis regeneration.

However, BLEU score is not linear with the length of the hypothesis, which makes the scoring process for each expanding step of hypothesis regeneration very slow. To further speed up the beam search procedure, we use an extension of a linear function of a Taylor approximation to the logarithm of corpus BLEU which was developed by (Tromble et al., 2008). The original BLEU score of two hypotheses e and e' are computed as follows.

$$BLEU(e, e') = \gamma(e, e') \times \exp\left(\frac{1}{4} \sum_{n=1}^{4} \log(P_n(e, e'))\right) \quad (3)$$

where $P_n(e, e')$ is the precision of n-grams in the hypothesis e given e' and $\gamma(e, e')$ is a brevity penalty. Let $|e|$ denote the length of e. The corpus log-BLEU gain is defined as follows:

$$\log(BLEU(e, e')) = \min\left(0, 1 - \frac{|e|}{|e'|}\right) + \frac{1}{4} \sum_{n=1}^{4} \log(P_n(e, e')) \quad (4)$$

Therefore, the first-order Taylor approximation to the logarithm of corpus BLEU is shown in Equation (5).

$$G(e, e') = \theta_0 |e| + \frac{1}{4} \sum_{n=1}^{4} \theta_n \cdot c_n(e, e') \quad (5)$$

where $c_n(e, e')$ are the counts of the matched n-grams and θ_n ($0 \le n \le 4$) are constant weights estimated with held-out data.

Suppose we have computed the expected n-gram counts from the N-best list or translation forest. Then we may extend linear corpus BLEU in (5) to n-gram expectation-based linear corpus BLEU to score the partial hypotheses h. That is

$$G(h, e') = \theta_0 |h| + \frac{1}{4} \sum_{n=1}^{4} \theta_n \cdot \sum_{t \in T_n} E[c_n(e', t)] \cdot \delta_n(h, t) \quad (6)$$

where $\delta_n(h, t)$ are n-gram indicator functions that equal 1 if n-gram t appears in h and 0 otherwise; $E[c_n(e', t)]$ ($1 \le n \le 4$) are the real-valued n-gram expectations. Different from lattice MBR decoding, n-gram expectations in this work are computed over the original translation N-best list or translation forest; T_n ($1 \le n \le 4$) are the sets of n-grams collected from translation N-best list or translation forest. Then we make a further extension: the expectations of the n-gram counts for each expanding step are computed over the partial translations. The lengths of all partial hypotheses are the same in each n-gram expanding step. For instance, in the 5th n-gram expanding step, the lengths of all the partial hypotheses are 5 words. Therefore, we use n-gram count expectations computed over partial original translations that only contain the first 5 words. The reason is that this solution contains more information about word orderings, since some n-grams appear more than others at the beginning of the translations while they may appear with the same or even lower frequencies than others in the full translations.

Once the expanding process of hypothesis regeneration is finished, we use a more precise BLEU metric to score all the translation candidates. We extend BLEU score in (3) to n-gram expectation-based BLEU. That is:

$$Score(h) = BLEU(h, e') = \exp\left[\min\left(0, 1 - \frac{E[|e'|]}{|h|}\right) + \frac{1}{4} \sum_{n=1}^{4} \log \frac{\sum_{t \in T_n} \min(c_n(h, t), E[c_n(e', t)])}{\sum_{t \in T_n} c_n(h, t)}\right] \quad (7)$$

where $c_n(h, t)$ is the count of n-gram t in the hypothesis h. The step of choosing the final translation is the same as fast consensus decoding (DeNero et al., 2009): first we compute n-

gram feature expectations, and then we choose the translation that is most similar to the others via expected similarity according to feature-based BLEU score as shown in (7). The difference is the space of translations: the space of fast consensus decoding is the same as MBR decoding, while the space of hypothesis regeneration is enlarged by the new translations produced via *n*-gram expansion.

2.3 Fast consensus hypothesis regeneration

We first generate two new hypothesis lists via forward and backward *n*-gram expansion using the scoring function in Equation (6). Then we choose a final translation using the scoring function in Equation (7) from the union of the original hypotheses and newly generated hypotheses. The original hypotheses are from the *N*-best list or extracted from the translation forest. The new hypotheses are generated by forward or backward *n*-gram expansion or are the union of both two new hypothesis lists (this is called "bi-directional *n*-gram expansion").

3 Experimental Results

We carried out experiments based on translation *N*-best lists generated by a state-of-the-art phrase-based statistical machine translation system, similar to (Koehn et al., 2007). In detail, the phrase table is derived from merged counts of symmetrized IBM2 and HMM alignments; the system has both lexicalized and distance-based distortion components (there is a 7-word distortion limit) and employs cube pruning (Huang and Chiang, 2007). The baseline is a log-linear feature combination that includes language models, the distortion components, translation model, phrase and word penalties. Weights on feature functions are found by lattice MERT (Macherey et al., 2008).

3.1 Data

We evaluated with different language pairs: Chinese-to-English, and German-to-English. Chinese-to-English tasks are based on training data for the NIST [1] 2009 evaluation Chinese-to-English track. All the allowed bilingual corpora have been used for estimating the translation model. We trained two language models: the first one is a 5-gram LM which is estimated on the target side of the parallel data. The second is a 5-gram LM trained on the so-called English *Gigaword corpus*.

			Chi	Eng
Parallel Train	Large Data	|S|	10.1M	
		|W|	270.0M	279.1M
Dev		|S|	1,506	1,506×4
Test	NIST06	|S|	1,664	1,664×4
	NIST08	|S|	1,357	1,357×4
Gigaword		|S|	-	11.7M

Table 1: Statistics of training, dev, and test sets for Chinese-to-English task.

We carried out experiments for translating Chinese to English. We first created a development set which used mainly data from the NIST 2005 test set, and also some balanced-genre web-text from the NIST training material. Evaluation was performed on the NIST 2006 and 2008 test sets. Table 1 gives figures for training, development and test corpora; |S| is the number of the sentences, and |W| is the size of running words. Four references are provided for all dev and test sets.

For German-to-English tasks, we used WMT 2006[2] data sets. The parallel training data contains about 1 million sentence pairs and includes 21 million target words; both the dev set and test set contain 2000 sentences; one reference is provided for each source input sentence. Only the target-language half of the parallel training data are used to train the language model in this task.

3.2 Results

Our evaluation metric is IBM BLEU (Papineni et al., 2002), which performs case-insensitive matching of *n*-grams up to $n = 4$.

Our first experiment was carried out over 1000-best lists on Chinese-to-English task. For comparison, we also conducted experiments with rescoring (two-pass) and three-pass hypothesis regeneration with only forward *n*-gram expansion as proposed in (Chen et al., 2008). In the "rescoring" and "three-pass" systems, we used the same rescoring model. There are 21 rescoring features in total, mainly translation lexicon scores from IBM and HMM models, posterior probabilities for words, *n*-grams, and sentence length, and language models, etc. For a complete description, please refer to (Ueffing et al., 2007). The results in BLEU-4 are reported in Table 2.

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

[2] http://www.statmt.org/wmt06/

testset	NIST'06	NIST'08
baseline	35.70	28.60
rescoring	36.01	28.97
three-pass	35.98	28.99
FCD	36.00	29.10
Fwd.	36.13	29.19
Bwd.	36.11	29.20
Bid.	36.20	29.28

Table 2: Translation performances in BLEU-4(%) over 1000-best lists for Chinese-to-English task: "rescoring" represents the results of rescoring; "three-pass", three-pass hypothesis regeneration with forward *n*-gram expansion; "FCD", fast consensus decoding; "Fwd", the results of hypothesis regeneration with forward *n*-gram expansion; "Bwd", backward *n*-gram expansion; and "Bid", bi-directional *n*-gram expansion.

Firstly, rescoring improved performance over the baseline by 0.3-0.4 BLEU point. Three-pass hypothesis regeneration with only forward *n*-gram expansion ("three-pass" in Table 2) obtained almost the same improvements as rescoring. Three-pass hypothesis regeneration exploits more hypotheses than rescoring, while rescoring involves more scoring feature functions than the former. They reached a balance in this experiment. Then, fast consensus decoding ("FCD" in Table 2) obtains 0.3-0.5 BLEU point improvements over the baseline. Both forward and backward *n*-gram expansion ("Fwd." and "Bwd." in Table 2) improved about 0.1 BLEU point over the results of consensus decoding. Fast consensus hypothesis regeneration (Fwd. and Bwd. in Table 2) got better improvements than three-pass hypothesis regeneration ("three-pass" in Table 2) by 0.1-0.2 BLEU point. Finally, combining hypothesis lists from forward and backward *n*-gram expansion ("Bid." in Table 2), further slight gains were obtained.

testset	Average time
three-pass	3h 54m
Fwd.	25m
Bwd.	28m
Bid.	40m

Table 3: Average processing time of NIST'06 and NIST'08 test sets used in different systems. Times include *n*-best list regeneration and re-ranking.

Moreover, fast consensus hypothesis regeneration is much faster than the three-pass one, because the former only needs to compute one feature, while the latter needs to compute more than 20 additional features. In this experiment, the former is about 10 times faster than the latter in terms of processing time, as shown in Table 3.

In our second experiment, we set the size of *N*-best list *N* equal to 10,000 for both Chinese-to-English and German-to-English tasks. The results are reported in Table 4. The same trend as in the first experiment can also be observed in this experiment. It is worth noticing that enlarging the size of the *N*-best list from 1000 to 10,000 did not change the performance significantly. Bi-directional n-gram expansion obtained improvements of 0.24 BLEU-score for WMT 2006 de-en test set; 0.55 for NIST 2006 test set; and 0.72 for NIST 2008 test set over the baseline.

Lang.	ch-en		de-en
testset	NIST'06	NIST'08	Test2006
baseline	35.70	28.60	26.92
FCD	36.03	29.08	27.03
Fwd.	36.16	29.25	27.11
Bwd.	36.17	29.22	27.12
Bid.	36.25	29.32	27.16

Table 4: Translation performances in BLEU-4 (%) over 10K-best lists.

We then tested the effect of the extension according to which the expectations over n-gram counts are computed on partial hypotheses rather than whole candidate translations as described in Section 2.2. As shown in Table 5, we got tiny improvements on both test sets by computing the expectations over n-gram counts on partial hypotheses.

testset	NIST'06	NIST'08
full	36.11	29.14
partial	36.13	29.19

Table 5: Translation performances in BLEU-4 (%) over 1000-best lists for Chinese-to-English task: "full" represents expectations over n-gram counts that are computed on whole hypotheses; "partial" represents expectations over n-gram counts that are computed on partial hypotheses.

3.3 Discussion

To speed up the search, the partial hypotheses in each expanding step are pruned. When pruning is applied, forward and backward *n*-gram expansion would generate different new hypothesis lists. Let us look back at the example in Figure 1.

Given 5 original hypotheses in Figure 1, if we set the beam size equal to 5 (the size of the original hypotheses), the forward and backward *n*-gram expansion generated different new hypothesis lists, as shown in Figure 2.

forward	backward
one week's work .	*one week's work .*
about week's work	*about one week's work*

Figure 2: Different new hypothesis lists generated by forward and backward *n*-gram expansion.

For bi-directional *n*-gram expansion, the chosen translation for a source sentence comes from the decoder 94% of the time for WMT 2006 test set, 90% for NIST test sets; it comes from forward n-gram expansion 2% of the time for WMT 2006 test set, 4% for NIST test sets; it comes from backward n-gram expansion 4% of the time for WMT 2006 test set, 6% for NIST test sets. This proves bidirectional *n*-gram expansion is a good way of enlarging the search space.

4 Conclusions and Future Work

We have proposed a fast consensus hypothesis regeneration approach for machine translation. It combines the advantages of feature-based consensus decoding and hypothesis regeneration. This approach is more efficient than previous work on hypothesis regeneration, and it explores a wider search space than consensus decoding, resulting in improved performance. Experiments showed consistent improvements across language pairs.

Instead of *N*-best lists, translation lattices or forests have been shown to be effective for MBR decoding (Zhang and Gildea, 2008; Tromble et al., 2008), and DeNero et al. (2009) showed how to compute expectations of *n*-grams from a translation forest. Therefore, our future work may involve hypothesis regeneration using an *n*-gram language model trained on the translation forest.

References

B. Chen, M. Federico and M. Cettolo. 2007. Better N-best Translations through Generative n-gram Language Models. In: *Proceedings of MT Summit XI*. Copenhagen, Denmark. September.

B. Chen, M. Zhang, A. Aw, and H. Li. 2008. Regenerating Hypotheses for Statistical Machine Translation. In: *Proceedings of COLING*. pp105-112. Manchester, UK, August.

J. DeNero, D. Chiang and K. Knight. 2009. Fast Consensus Decoding over Translation Forests. In: *Proceedings of ACL*. Singapore, August.

J. Duchateau, K. Demuynck, and P. Wambacq. 2001. Confidence scoring based on backward language models. In: *Proceedings of ICASSP 2001*. Salt Lake City, Utah, USA, May.

L. Huang and D. Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In: *Proceedings of ACL*. pp. 144-151, Prague, Czech Republic, June.

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*. pp. 177-180, Prague, Czech Republic.

S. Kumar and W. Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In: *Proceedings of NAACL*. Boston, MA, May.

W. Macherey, F. Och, I. Thayer, and J. Uszkoreit. 2008. Lattice-based Minimum Error Rate Training for Statistical Machine Translation. In: *Proceedings of EMNLP*. pp. 725-734, Honolulu, USA, October.

F. Och. 2003. Minimum error rate training in statistical machine translation. In: *Proceedings of ACL*. Sapporo, Japan. July.

F. 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: *Proceedings of NAACL*. Boston.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the ACL 2002*.

R. Tromble, S. Kumar, F. J. Och, and W. Macherey. 2008. Lattice minimum Bayes-risk decoding for statistical machine translation. In: *Proceedings of EMNLP*. Hawaii, US. October.

N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. 2007. NRC's Portage system for WMT 2007. In: *Proceedings of ACL Workshop on SMT*. Prague, Czech Republic, June.

H. Zhang and D. Gildea. 2008. Efficient multipass decoding for synchronous context free grammars. In: *Proceedings of ACL*. Columbus, US. June.