Hybrid Decoding: Decoding with Partial Hypotheses Combination over Multiple SMT Systems*

Lei Cui[†], **Dongdong Zhang**[‡], **Mu Li**[‡], **Ming Zhou**[‡], and **Tiejun Zhao**[†]

[†]School of Computer Science and Technology
Harbin Institute of Technology
{cuilei,tjzhao}@mtlab.hit.edu.cn
[‡]Microsoft Research Asia
{dozhang,muli,mingzhou}@microsoft.com

Abstract

In this paper, we present *hybrid decoding* — a novel statistical machine translation (SMT) decoding paradigm using multiple SMT systems. In our work, in addition to component SMT systems, system combination method is also employed in generating partial translation hypotheses throughout the decoding process, in which smaller hypotheses generated by each component decoder and hypotheses combination are used in the following decoding steps to generate larger hypotheses. Experimental results on NIST evaluation data sets for Chinese-to-English machine translation (MT) task show that our method can not only achieve significant improvements over individual decoders, but also bring substantial gains compared with a state-of-the-art word-level system combination method.

1 Introduction

In recent years, system combination for SMT has been known to be quite effective with translation consensus information built from multiple SMT systems. The combination approaches can be classified into two types. One is the combination with each system's outputs, which can be seen as full hypotheses combination. The other is the partial hypotheses (PHS) combination during the decoding phase.

A lot of impressive work has been done to improve the performance of the SMT systems by utilizing consensus statistics which come from single system or multiple systems. For example, Minimum Bayes Risk (MBR) (Kumar and Byrne, 2004) decoding over n-best list finds a translation that has lowest expected loss with all the other hypotheses, and it shows that improvement over the Maximum a Posteriori (MAP) decoding. Several word-based methods (Rosti et al., 2007a; Sim et al., 2007) have also been proposed. Usually, these methods take n-best list from different SMT systems as inputs, and construct a confusion network for second-pass decoding. There are also a lot of research work to advance the confusion network construction by finding better alignment between the skeleton and the other hypotheses (He et al., 2008; Ayan et al., 2008). Typically, all the approaches above only use full hypotheses but have no access to the PHS information.

Moreover, some dedicated efforts have been tried by manipulating PHS between multiple MT systems. Collaborative decoding (co-decoding) (Li et al., 2009) leverages translation consensus by exchanging partial translation results and reranking both full and partial hypotheses explored in decoding. However, no new PHS are generated compared to the individual decoding but only the ranking is affected. Liu et al. (2009) proposes joint decoding, a method that integrates multiple translation models in one decoder. Although joint decoding is able to generate new translations compared to single decoder, it has to use the PHS existed in one of its component decoder at each step. Different from their work, we propose a new perspective which leverages outputs from local word-level combination. This will potentially bring much benefit of performance since word-

*This work has been done while the first author was visiting Microsoft Research Asia.

Coling 2010: Poster Volume, pages 214–222,
Beijing, August 2010

level combination can produce more promising PHS.

The word-level system combination method is employed to generate partial translation hypotheses in our hybrid decoding framework. In this sense, full hypotheses word-level combination (FH-Comb) method(Rosti et al., 2007a; Sim et al., 2007; He et al., 2008; Ayan et al., 2008) can be considered as a special case of hybrid decoding, where their combinations are only performed on the largest hypotheses. Similar with FH-Comb, hybrid decoding also uses word alignment information. However, challenge exists in hybrid decoding as word alignment needs to be carefully conducted through the decoding process. Obviously, document-level word alignment methods such as GIZA++(Och and Ney, 2000) are quite time consuming and unpractical to be embedded into hybrid decoding. We propose a heuristic method that can conduct word alignment of partial hypotheses based on word alignment information of phrase pairs learnt automatically from the model training process. In this way, more PHS are generated and the search space is enlarged substantially, which brings better translation results.

The rest of the paper is organized as follows: Section 2 gives a formal description of hybrid decoding, including framework overview, word-level PHS combination and parameter estimation. We conduct experiments with different settings and make comparison between our method and baseline, as well as a state-of-the-art word-level system combination method in Section 3. Experimental results discussion is presented in Section 4. Section 5 concludes the paper.

2 Hybrid Decoding

2.1 Overview

Different system combination methods (Li et al., 2009; Liu et al., 2009) offer different frameworks to coordinate multiple SMT decoders. Hybrid decoding provides a new scheme to organize multiple decoders to work synchronously. As the decoding algorithms may differ in multiple decoders[1], hybrid decoding has some difficulty in integrating different decoding algorithms. Without loss of generality, we assume that bottom-up CKY-based decoding is adopted in each individual decoder, which is the same as co-decoding (Li et al., 2009) and joint decoding (Liu et al., 2009). Hybrid decoding collects n-best PHS of a source span[2] from multiple decoders, then results from word-level PHS combination of that span are given back to each decoder, mixed with the original PHS. After that, we re-rank the hybrid list and continue the decoding. In an example with two decoders, parts of the whole decoding process are illustrated in Figure 1 and can be summarized as follows:

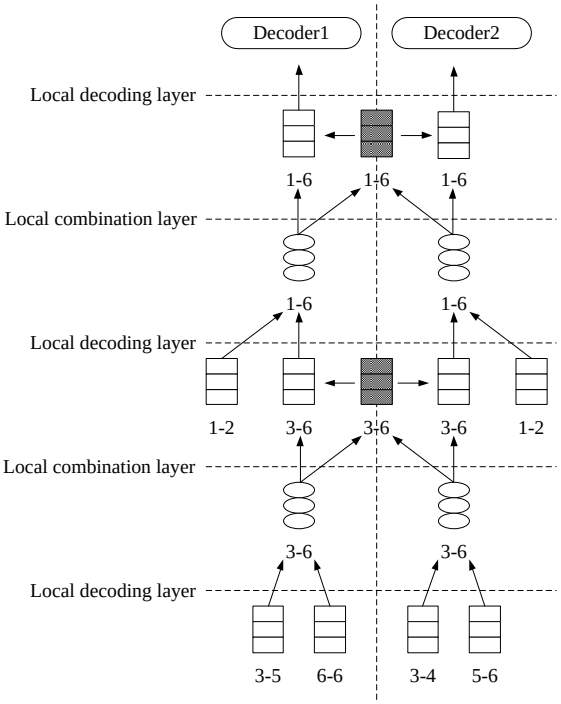

Figure 1: Hybrid decoding with two decoders, where the string "s-e" means the source span starts from position s and ends at position e. The blank rectangles represent the n-best partial translations of each decoder, and the shaded rectangles illustrate the n-best local combination outputs. The ovals denote bottom-up CKY-based decoding results.

[1]In the SMT area, some decoders use left-right decoding to generate the hypothesis and "Pharaoh" (Koehn et al., 2003) is one of them, while others adopt bottom-up decoding which is represented by "Hiero" (Chiang, 2007).

[2]The word "span" is used to represent translation unit in CKY-based decoders, which denotes one or more consecutive words in the source sentence.

1. *Individual decoding*. Each individual decoder should maintain the n-best PHS of each span from the bottom. After all the individual decoders finish translating the same span, they feed their own partial translations into a public container which can be used for word-level PHS combination, then get back the partial combination outputs for step 3.

2. *Local word-level combination*. After fed with PHS from multiple decoders, a confusion network is built and word-level combination for PHS is conducted. The obtained new partial translations are given back to each individual decoder to continue the decoding.

3. *Mix new PHS with the original ones*. The span in each individual decoder will receive the corresponding new PHS from the local combination outputs. The feature space of the new PHS is not exactly the same with that of the original ones. It has to be mapped in some way then the mixed hypotheses are reranked.

In the following sub-sections, we first present the background of word-level combination for PHS, then introduce hybrid decoding algorithm in detail, as well as the feature definition and parameter estimation.

2.2 Word-Level Combination for Partial Hypotheses

Most word-level system combination methods are based on confusion network decoding. In confusion network construction, one hypothesis has to be selected as the skeleton which determines the word order of the combination results. Other hypotheses are aligned against the skeleton. Either votes or some word confidence scores are assigned to each word in the network.

Most of the research on confusion network construction focuses on seeking better word alignment between the skeleton and the other hypotheses. So far, several word alignment procedures are used for SMT system combination, which mainly are GIZA++ alignments (Matusov et al., 2006), TER alignments (Sim et al., 2007) and IHMM alignments (He et al., 2008). Similar with general word-level system combination method, word-level PHS combination also uses word alignment information. However, in hybrid decoding, it is quite time-consuming and impractical to conduct word alignment like GIZA++ for each span. Fortunately, unit hypotheses word alignment can be obtained from the model training process, which is shown in Figure 2. We devise a heuristic approach for PHS alignment that leverages the translation derivations from the sub-phrases. The derivation information ultimately comes from the phrase table in phrase-based systems (Koehn et al., 2003; Xiong et al., 2006) or the rule table in syntactic-based systems (Chiang, 2007; Liu et al., 2007; Galley et al., 2006).

政治||| political ||| 0-0
政治 经济||| political and economic ||| 0-0 1-2
经济||| economic ||| 0-0
经济 利益||| economic interests ||| 0-0 1-1
政治 [X_1] ||| political and [X_1] ||| 0-0 1-2

Figure 2: The example of translation alignment from phrase-table and rule-table

The derivation is built in a phrase-based system as follows. For example, we have two phrase translations “我们 的 ||| our ||| 0-0 1-0” and “经济 利益 ||| economic interests ||| 0-0 1-1”, where string “m-n” means the m^{th} word in the source phrase is aligned to the n^{th} word in the target phrase. When combining the two phrases for generating “我们 的 经济 利益”, we obtain the translation hypothesis as “our economic interests” and also integrate the alignment fragment to get “0-0 1-0 2-1 3-2”. The case is similar in syntactic-based system for non-terminal substitution, which we will not discuss further here.

Next, we introduce the skeleton-to-hypothesis word alignment algorithm in detail. With the translation derivations, the skeleton-to-hypothesis ($sk2hy$) word alignment can be performed based on the source-to-skeleton ($so2sk$) and source-to-hypothesis ($so2hy$) word alignment as they share the same source sentence. The basic idea is to construct the $sk2hy$ word alignment with the *minimum correspondence subsets* (MCS). A MCS is defined as a triple $< SK, HY, SO >$ where the

SK is the subset of skeleton words, HY is the subset of the hypothesis words, and SO is the minimum source word set that all target words in both SK and HY are aligned to. Figure 3 shows the algorithm for skeleton-to-hypothesis alignment. Most of the pseudo-code is self-explained except for some subroutines, which are listed in Table 1.

```
1:  procedure SKEHYPALIGN(so2sk, so2hy)
2:      repeat
3:          Fetch out a source word to SO
4:          SO_1 = SO_2 = SO
5:          repeat
6:              SO=UNION(SO_1, SO_2)
7:              SK=GETALIGN(SO, so2sk)
8:              HY=GETALIGN(SO, so2hy)
9:              SO_1=GETALIGN(SK, so2sk)
10:             SO_2=GETALIGN(HY, so2hy)
11:         until |SO_1| == |SO_2| == |SO|
12:         sim_max = -infinity
13:         for all sk ∈ SK do
14:             for all hy ∈ HY do
15:                 sim =SIM(sk, hy)
16:                 if sim ≥ sim_max then
17:                     sim_max = sim
18:                     sk_max = sk
19:                     hy_max = hy
20:                 end if
21:             end for
22:         end for
23:         ADDALIGN(sk_max, hy_max)
24:     until all the source words are fetched out
25: end procedure
```

Figure 3: Algorithm for skeleton-to-hypothesis alignment

Subroutines	Description
UNION(A,B)	the union of set A and set B
GETALIGN(S,$align$)	get the words aligned to S based on $align$
SIM(w_1,w_2)	similarity between w_1 and w_2, we use edit distance here
ADDALIGN(w_1,w_2)	align w_1 with w_2

Table 1: Description for subroutines

Due to the variety of the word order in n-best outputs, skeleton selection becomes essential in confusion network construction. The simplest way is to use the top-1 PHS from any individual decoder with the best performance under some criteria. However, this cannot always lead to better performance on some evaluation metrics (Rosti et al., 2007a). An alternative would be MBR method with some loss function such as TER (Snover et al., 2006) or BLEU (Papineni et al., 2002). We show the experimental results of two skeleton selection methods for PHS combination in Section 3.

2.3 Hybrid Decoding Model

For a given source sentence f, any individual decoder in hybrid decoding finds the best translation e^* among the possible translation hypotheses $\Phi(f)$ in terms of a ranking function F:

$$e^* = \text{argmax}_{e \in \Phi(f)} F(e) \quad (1)$$

Suppose we have n individual decoders. The ranking function F_n of the n^{th} decoder can be written as:

$$F_n(e) = \sum_{i=1}^{m} \lambda_{n,i} h_{n,i}(f, e) \quad (2)$$

where each $h_{n,i}(f, e)$ is a feature function of the n^{th} decoder, and $\lambda_{n,i}$ is the corresponding feature weight. m is the number of features in each decoder.

The final result of hybrid decoder is the top-1 translation from the confusion network, which is constructed on multiple decoders with the last layer's output of CKY-based decoding.

2.4 Hybrid Decoding Algorithm

The hybrid decoder acts as a control unit which controls the synchronization of multiple individual decoders. The algorithm is fully demonstrated in Figure 4. The hybrid decoder pushes the same span f_i^j to different decoders and gets back the n-best PHS (lines 2-6). When the span's length is too small, both word alignment and partial combination results are not accurate. We predefine a fixed threshold δ which is used for determining the start-up of combination (line 7). When the length condition holds, the n-best PHS of each individual

decoder are stored in container G (lines 8). Confusion network is constructed and new PHS can be extracted from it and are further mixed and sorted with the original ones (lines 11-15).

```
1:  procedure HYBRIDDECODING(f_1^n, D)
2:      for l ← 1...n do
3:          for all i, j s.t. j − i = l do
4:              G ← ∅
5:              for all d ∈ D do
6:                  nbest =DECODING(d, i, j)
7:                  if j − i ≥ δ then
8:                      ADD(G, nbest)
9:                  end if
10:             end for
11:             cn =CONNETBUILD(G)
12:             nbest′ =GETPARHYP(cn)
13:             for all d ∈ D do
14:                 MIXSORT(nbest_d, nbest′)
15:             end for
16:         end for
17:     end for
18: end procedure
```

Figure 4: Hybrid decoding algorithm

2.5 Hybrid Decoding Features

Next we present the PHS word-level combination feature functions for hybrid decoding. Following (Rosti et al., 2007b), four features are utilized to model the PHS as:

Word Confidence Feature $h_{wc}(e)$
The word confidence feature is computed as $h_{wc}(e) = \sum_{i=1}^{n} \mu_i c_{iw}$, where n is the number of the systems, μ_i is the system confidence of system i, and c_{iw} is the word confidence of word w in system i.

Word Penalty Feature $h_{wp}(e)$
Word penalty feature is the number of words in the partial hypothesis (PH).

Null Penalty Feature $h_{np}(e)$
For null penalty feature, we mean the number of NULL links along the PH when extracted from the confusion network.

Language Model Feature $h_{lm}(e)$
Different from the above three combination features, which can be obtained during the confusion network construction or hypotheses extraction, the language model feature cannot be summed up on the fly. Instead, it must be re-computed when building each new PH.

2.6 Feature Space Mapping

The features used in hybrid decoding can be classified into two categories: features for individual decoders (FID) and features for PHS word-level combination (FComb), and they are independent. When mixing the new PHS with the original ones of individual decoders, FComb space has to be mapped to a FID space. However, several features in FID are not defined in FComb, such as source to target (S2T) phrase probability, target to source (T2S) phrase probability, S2T lexical probability, T2S lexical probability and other model specific features. A mapping function H needs to be defined as follows:

$$F_{fid} = H(F_{fcomb}) \tag{3}$$

where F_{fcomb} denotes the feature vector from FComb space, while F_{fid} is the feature vector from FID space.

An easy mapping function is implemented with an intuitive motivation: PHS combination results are better than the ones in individual decoder and we prefer not to disorder the original search space. Thus, the undefined feature values of PHS from FComb space are assigned by corresponding feature values of the top-1 PH in original decoder. Experiments show that our method is not only practical but also quite effective.

2.7 Parameter Estimation

Minimum Error Rate Training (MERT) (Och, 2003) algorithm is adopted to estimate feature weights for hybrid decoding. As hybrid decoder makes use of PHS from both individual decoders and combination results as a whole, we devise a new feature vector representation. The feature vectors from FID space and FComb space are simply concatenated to form a longer vector without overlapping. The weights are tuned simultaneously in order to reach a relatively global optima.

3 Experiment

3.1 Data and Metric

We conducted our experiments on the test data of NIST 2005 and NIST 2006 Chinese-to-English machine translation tasks. The NIST 2003 test data is used as the development data to tune the parameters. Statistics of the data sets are shown in Table 2. Translation performances are measured with case-insensitive BLEU4 score (Papineni et al., 2002). Statistical significance test is performed using the bootstrap re-sampling method proposed by Koehn (2004).

The bilingual training corpora we used are listed in Table 3, which contains 498K sentence pairs, 12.1M Chinese words and 13.8M English words after pre-processing. Word alignment is performed by GIZA++ (Och and Ney, 2000) in both directions with the default setting, and the intersect-diag-grow method is used to refine the symmetric word alignment.

Data Set	# Sentences
NIST 2003(dev)	919
NIST 2005(test)	1,082
NIST 2006(test)	1,664

Table 2: Statistics of test/dev data sets

LDC ID	Description
LDC2003E07	Ch/En Treebank Par Corpus
LDC2003E14	FBIS Multilanguage Texts
LDC2005T06	Ch News Translation Text Part 1
LDC2005T10	Ch/En News Magazine Par Text
LDC2005E83	GALE Y1 Q1 Release - Translations
LDC2006E26	GALE Y1 - En/Ch Par Financial News
LDC2006E34	GALE Y1 Q2 Release - Translations V2.0
LDC2006E85	GALE Y1 Q3 Release - Translations
LDC2006E92	GALE Y1 Q4 Release - Translations

Table 3: Training corpora for Chinese-English translation

The language model used for hybrid decoding and all the baseline systems is a 5-gram model trained with the Xinhua portion of LDC English Gigaword Version 3.0 plus the English part of bilingual training data.

3.2 Implementation

We use two baseline systems. The first one (SYS1) is re-implementation of Hiero, a hierarchical phrase-based system (Chiang, 2007) based on Synchronous Context Free Grammar (SCFG). Phrasal translation rules and hierarchical translation rules with nonterminals are extracted from all the bilingual sentence pairs. The second one (SYS2) is a phrase-based system (Xiong et al., 2006) based on Bracketing Transduction Grammar (Wu, 1997) with a lexicalized reordering model (Zhang et al., 2007) under maximum entropy framework, where the phrasal translation rules are exactly the same with that of SYS1. The lexicalized reordering model is trained using the MaxEnt toolkit (Zhang, 2006) where the training instances are extracted from subset of the training corpora, which contains LDC2003E07, LDC2003E14, LDC2005T06, LDC2005T10. Both systems use the bottom-up CKY-based decoding with cube-pruning (Chiang, 2007) and the beam size is set to 10 for decoding efficiency.

For hybrid decoder, we set δ to be $sentence.length - 3$, meaning that the PHS of individual decoders only perform local combination in the last three layers. The reason why we adopt this setting is because we find that starting local combination on short spans hurts the performance badly on test data. Experimental results are shown in the next section.

3.3 Translation Results

We present the overall results of hybrid decoding over two baseline systems on both test sets. We also implement an IHMM-based word-level system combination method (He et al., 2008) to make comparison with hybrid decoding system, and the n-best candidates used for IHMM-based word-level system combination is set to 10. Parameters for all the systems are tuned on NIST 2003 test set. The results are shown in Table 4.

In Table 4, we find that the hybrid decoding performs significantly better than SYS1 and SY2 on both test sets. Besides, compared to IHMM word-level system combination method, hybrid decoding also brings substantial gains with 0.63% and 0.92% points respectively.

	NIST 2005	NIST 2006
SYS1	0.3745	0.3346
SYS2	0.3699	0.3296
IHMM Word-Comb	0.3821^{*}	0.3421^{*}
Hybrid	0.3884^{*+}	0.3513^{*+}

Table 4: Hybrid decoding results on test sets, *:significantly better than SYS1 and SYS2 with $\boldsymbol{p}<0.01$, +:significantly better than IHMM Word-Comb with $\boldsymbol{p}<0.01$

We also try different layers for determining the start-up of local word-level PHS combination. Figure 5 gives the intuitive BLEU results.

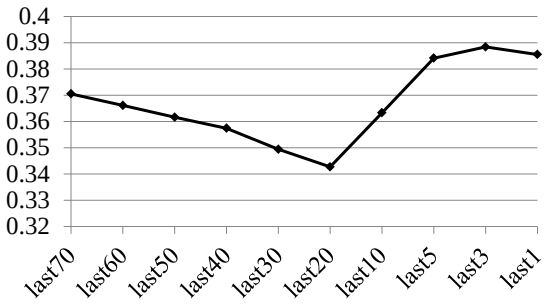

Figure 5: Performance of hybrid decoding with different start-up settings on NIST 2005 test set, where the "lastM" means to conduct local word-level PHS combination in the last M layers from the perspective of CKY-based decoding.

As shown in Figure 5, the performance drops drastically if we start to conduct word-level PHS combination too early. After considering about efficiency and performance, we determine to do that in the last three layers.

We then investigate the effects on hybrid decoding with different beam sizes, and compare the trend with two baseline systems and IHMM-based word-level system combination method as well. The results are illustrated in Figure 6.

From what we see in Figure 6, the performance of each system is monotonically increasing as the beam size becomes larger. Hybrid decoding performs consistently better than IHMM Word-Comb when the beam size is small, and the largest improvement (+0.63% points) is obtained when the beam size is set to 10. However, as the beam size

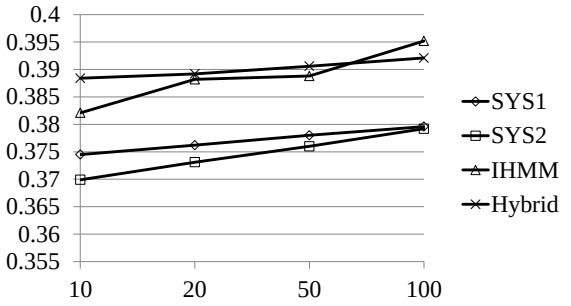

Figure 6: Performance of hybrid decoding with different beam sizes on NIST 2005 test set

increases, the performance gap is getting narrow. One intuitive observation is that hybrid decoding performs slightly worse than IHMM Word-Comb when the beam size is set to 100. One possible reason for this phenomenon is that, the alignment noise may be introduced to hybrid decoding since we have to generate monolingual alignments with many poor translation derivations.

The confusion network for PHS of each system can be built independently. We would like to evaluate the performance of single system hybrid decoding. Table 5 gives the results on both Hiero and BTG decoders.

	NIST 2005		NIST 2006	
	SYS1	SYS2	SYS1	SYS2
baseline	0.3745	0.3699	0.3346	0.3296
self-comb	0.3770	0.3758^{*}	0.3358	0.3355^{*}

Table 5: Hybrid decoding for single system, *:significantly better than baseline with $\boldsymbol{p}<0.05$

Table 5 shows that BTG decoder (SYS2) has more potential for so-called "self-boosting". The self-combination of BTG decoder improves the performance substantially over the baseline. However, we did not observe any significant improvement for Hiero decoder (SYS1).

Finally, we examine the impacts of skeleton selection for PHS in hybrid decoding. The results in Table 6 demonstrate that, compared to the top-1 selection method, translation performance can be improved significantly with MBR-based skeleton selection method. It strongly suggests that choosing the skeleton with more consistent word order

will lead to better translation results.

	NIST 2005	NIST 2006
Top-1	0.3817	0.3415
MBR	0.3884*	0.3513*

Table 6: Skeleton selection in hybrid decoding, *:significantly better than top-1 skeleton selection method with $p<0.01$

4 Discussion

System combination methods have been widely used in SMT to improve the performance. For example, in (Rosti et al., 2007a), several combination methods have been proposed to make use of different kinds of consensus information. In (He et al., 2008), better word alignment method is adopted to advance the word-level system combination. Our method is different from these methods in the sense that we do not exclusively rely on the n-best full hypotheses from each individual decoder, but emphasize the importance of word-level combination for PHS. Thus, it enlarges the search space and is more prone to find better translations. Experimental results have shown the effectiveness of our method.

The idea of multiple systems collaborative decoding (Li et al., 2009) works well on re-ranking the outputs of each system using n-gram agreement statistics. However, no new translation results are generated compared to individual decoding. Our method takes advantage of confusion network to give PHS which cannot be seen before.

Although (Liu et al., 2009) also work on PHS, we explore the cooperation of multiple systems from a new perspective. They use translation derivations from different decoders jointly as a bridge to connect different models. Different from their work, we devise a heuristic method to obtain word alignment information from the derivation of each decoder, which can be embedded for word-level PHS combination easily and efficiently.

5 Conclusion and Future Work

In this paper, we propose a new SMT decoding framework named hybrid decoding, in which multiple decoders work synchronously to conduct local decoding and local word-level PHS combination in turn. We also devise a heuristic method to obtain word alignment information directly from the translation derivations, which is both intuitive and efficient. Experimental results show that with hybrid decoding the overall performance can be improved significantly over both the individual baseline decoder and the state-of-the-art system combination method.

In the future, we will involve more individual SMT decoders into hybrid decoding. In addition, we would like to keep on this work in two directions. On the one hand, start-up threshold of PHS combination will be explored in detail to find its underlying impact on hybrid decoding. On the other hand, we will try to employ a more theoretically sound approach to conduct the feature space mapping from the feature space of confusion network to that of individual decoders.

References

Ayan, Necip Fazil, Jing Zheng, and Wen Wang. 2008. *Improving alignments for better confusion networks for combining machine translation systems*. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 33-40

Chiang, David. 2005. *A hierarchical phrase-based model for statistical machine translation*. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263-270

Chiang, David. 2007. *Hierarchical phrase-based translation*. *Computational Linguistics*, 33(2): pages 201-228

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. *Scalable inference and training of context-rich syntactic translation models*. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 961-968

He, Xiaodong, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. *Indirect-hmm-based hypothesis for combining outputs from machine translation systems*. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98-107

Koehn, Phillip, Franz J. Och, and Daniel Marcu. 2003. *Statistical phrase-based translation*. In *Proceed-*

ings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 48-54

Koehn, Phillip. 2004. *Statistical significance tests for machine translation evaluation*. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388-395

Kumar, Shankar and William Byrne. 2004. *Minimum bayes-risk decoding for statistical machine translation*. In *Proceedings of the 2004 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 169-176

Li, Mu, Nan Duan, Dongdong Zhang, Chi-Ho Li, and Ming Zhou. 2009. *Collaborative decoding: partial hypothesis re-ranking using translation consensus between decoders*. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 585-592

Liu, Yang, Yun Huang, Qun Liu, and Shouxun Lin. 2007. *Forest-to-string statistical translation rules*. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 704-711

Liu, Yang, Haitao Mi, Yang Feng, and Qun Liu. 2009. *Joint decoding with multiple translation models*. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 576-584

Matusov, Evgeny, Nicola Ueffing, and Hermann Ney. 2006. *Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment*. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33-40

Och, Franz Josef. and Hermann Ney. 2000. *Improved statistical alignment models*. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440-447

Och, Franz Josef. 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160-167

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318

Rosti, Antti-Veikko, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. *Combining outputs from multiple machine translation systems*. In *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228-235

Rosti, Antti-Veikko, Spyros Matsoukas, and Richard Schwartz. 2007. *Improved word-level system combination for machine translation*. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312-319

Sim, K.C., W. Byrne, M. Gales, H. Sahbi, and P. Woodland. 2007. *Consensus network decoding for statistical machine translation combinnation*. In *32nd IEEE International Conference on Acoustics, Speech, and Signal Processing*

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciula, and John Makhoul. 2006. *A study of translation edit rate with targeted human annotation*. In *the 7th conference of the Association for Machine Translation in the Americas*, pages 223-231

Wu, Dekai. 1997. *Stochastic inversion transduction grammars and bilingual parsing of parallel corpora*. *Computational Linguistics*, 23(3): pages 377-404

Xiong, Deyi, Qun Liu, and Shouxun Lin. 2006. *Maximum entropy based phrase reordering model for statistical machine translation*. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 521-528

Zhang, Dongdong, Mu Li, Chi-Ho Li, Ming Zhou. 2007. *Phrase Reordering Model Integrating Syntactic Knowledge for SMT*. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 533-540

Zhang, Le. 2006. *Maximum entropy modeling toolkit for python and c++*. available at `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.