An Incremental Three-pass System Combination Framework by Combining Multiple Hypothesis Alignment Methods

Jinhua Du[1], Andy Way[1,2]
[1] Centre for Next Generation Localisation, Dublin City University, Dublin, Ireland
[2] National Centre for Language Technology, Dublin City University, Dublin, Ireland
{jdu, away}@computing.dcu.ie

Abstract

System combination has been applied successfully to various machine translation tasks in recent years. As is known, the hypothesis alignment method is a critical factor for the translation quality of system combination. To date, many effective hypothesis alignment metrics have been proposed and applied to the system combination, such as TER, HMM, ITER, IHMM, and SSCI. In addition, Minimum Bayes-risk (MBR) decoding and confusion networks (CN) have become state-of-the-art techniques in system combination. In this paper, we examine different hypothesis alignment approaches and investigate how much the hypothesis alignment results impact on system combination, and finally present a three-pass system combination strategy that can combine hypothesis alignment results derived from multiple alignment metrics to generate a better translation. Firstly, these different alignment metrics are carried out to align the backbone and hypotheses, and the individual CNs are built corresponding to each set of alignment results; then we construct a 'super network' by merging the multiple metric-based CNs to generate a consensus output. Finally a modified MBR network approach is employed to find the best overall translation. Our proposed strategy outperforms the best single confusion network as well as the best single system in our experiments on the NIST Chinese-to-English test set and the WMT2009 English-to-French system combination shared test set.

Keywords

system combination; three-pass system combination; hypothesis alignment; MBR decoder; confusion network; super network.

1 Introduction

In the past several years, multiple system combination has been shown to be helpful in improving translation quality. Recently, confusion network-based (CN-based) networks in (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007a; He et al., 2008), have become the state-of-the-art methodology to implement the combination strategy. A CN is essentially a directed acyclic graph which is built by a set of translation hypotheses against a reference or "backbone". Each arc between two nodes in the CN denotes a word or token, possibly a *null* item, with an associated posterior probability. Generally, like the translation decoding process in phrase-based statistical machine translation (PB-SMT), the CN decoding process also uses a log-linear model (Och and Ney, 2002), which combines a set of different features to search for the best path or an *N*-best list by dynamic programming algorithms.

Typically, the dominant CN is constructed on the word level by a state-of-the-art framework. Firstly, a minimum Bayes-risk (MBR) decoder (Kumar and Byrne, 2004) is utilised to choose the backbone from a merged set of hypotheses, and then the remaining hypotheses are aligned against the backbone by a specific alignment approach. Currently, most research in system combination has focussed on the hypothesis alignment due to its significant influence on combination quality. A TER-based (Snover et al., 2006) system combination strategy was first introduced in (Sim et al., 2007). More recently, many hypothesis alignment metrics have been proposed and successfully applied in system combination, such as IHMM (He et al., 2008) and ITG (Karakos et al., 2008). In all these papers, the proposed alignment method outperformed the TER-based baseline system.

A multiple CN or 'super network' framework was first proposed in (Rosti et al., 2007b), where the final CN was comprised of all individual system CNs—all constructed based on the same alignment metric, namely TER—as the backbone. In this method, the MBR decoder was not used so that the risk of selecting a poorly performing backbone was lessened. However, the potential problem is that from an engineering viewpoint the complexity increases where there are a lot of individual system results. A consensus network MBR (ConMBR) approach employs an MBR decoding to select the best one with the minimum cost from the original single system outputs compared to the consensus output (Sim et al., 2007). In this paper, we present a revised, extended idea proposed first in (Du and Way, 2009c) that employs the MBR, super network and a modified ConMBR to construct a three-pass system combination framework which can effectively combine different hypothesis alignment results and easily be extended to more alignment metrics. We demonstrate using two language pairs that such a framework is consistently effective when a number of different hypothesis alignment methods are combined.

The remainder of this paper is organised as follows. In section 2, we examine the impact of different hypothesis alignment methods on the performance of CN system combination. In section 3, we summarize four commonly used hypothesis alignment metrics in our combination task: TER, HMM, IHMM and SSCI (Du et al., 2009b), which have different working mechanisms and represent the metrics currently most preferred in system combination tasks. Section 4 introduces the modified ConMBR (mConMBR) decoding. Then, Section 5 describes the implementation details of our proposed three-pass combination strategy which can combine multiple different hypothesis alignment metrics. The experiments conducted on NIST Chinese-to-English and WMT2009 English-to-French data are reported in Sections 6 and 7. Section 8 concludes and gives avenues for future work.

2 The Impact of Hypothesis Alignment on the Confusion Network

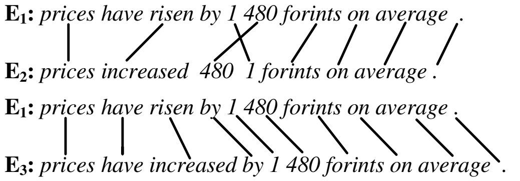
The process of hypothesis alignment is similar to the word alignment between source and target languages in PB-SMT (Och and Ney, 2003). The differences are firstly that the source and target sides are the same language, and secondly, the word alignment types are limited to 1-to-1, 1-to- *null* and *null*-to-1. Currently, the CN has two crucial characteristics: (i) it is a word-level graph; and (ii) a monotone decoding process is selected. Therefore, hypothesis alignment plays a vital role in the CN because the backbone sentence decides the skeleton and word order of the consensus output.

E$_1$: *prices have risen by 1 480 forints on average .*

E$_2$: *prices increased 480 1 forints on average .*

E$_3$: *prices have increased by 1 480 forints on average .*

(a) Hypotheses Set


E1: prices have risen by 1 480 forints on average .
E2: prices increased 480 1 forints on average .
E1: prices have risen by 1 480 forints on average .
E3: prices have increased by 1 480 forints on average .

(b) backbone-hypothesis alignment

E$_1$:	*prices*	*have*	*risen*	*by*	*1*	*480*	*forints*	*on*	*average*	.
E$_2$:	*prices*	@	*increased*	@	*1*	*480*	*forints*	*on*	*average*	.
E$_3$:	*prices*	*have*	*increased*	*by*	*1*	*480*	*forints*	*on*	*average*	.

(c) Normalise hypothesis alignment and construct confusion network

Figure 1. Normalise hypothesis alignment and construct confusion network

Figure 1 gives an overall view of the main steps involved in CN-based system combination, including how to align the hypotheses, carry out the word re-ordering as well as construct the CN. In Figure 1(a), hypotheses from different MT systems are merged to form a new *N*-best list, from which the backbone is selected using the MBR decoder. The most frequently used loss functions in MBR are TER and BLEU (Papineni et al., 2002). Then, as illustrated in Figure 1(b), assuming E_1 to be the selected backbone under some loss function, the remaining hypotheses are aligned against E_1 by carrying out a specific alignment metric as described in Section 3. The symbol @ denotes a *null* word. Note that there are only three types of word alignment in system combination, namely, 1-to-1, 1-to-*null* and *null*-to-1 in terms of bidirectional alignment. Depending on the particular method of word alignment, word reordering is carried out and a CN is constructed based on the reordered hypotheses as Figure 1 (c) shows. Finally, sets of local and global features are integrated into a log-linear model to decode the CN.

The most challenging problem for CN decoding is the phenomenon of "non-grammatical" phrases, which are mainly caused by the arbitrary word reordering and decoding strategy inside the CN. There might be several arcs between any two adjacent nodes. Each arc indicates an alternative word or *null.* The aim of the search process is to produce a sequence with the best overall score, while at each position, the selected word is mainly decided by methods such as voting, confidence scores, or relative probability of the candidates. Thus there may be no direct grammatical relationship between any adjacent words in the voting decision, as there is no guarantee that consecutive words in the output consensus translation come from the same CN. Although nowadays most MT research introduces some syntax-like features into the CN (such as a language model, for instance), it still cannot avoid producing "non-grammatical" output. However, a high-quality hypothesis alignment can reduce this kind of influence to some extent, since the more accurately the words are aligned, the less noise is produced.

When we examine the impact of hypothesis alignment on the CN, two key issues should be studied. The first one is that of word order: how does the word order impact on the skeleton of the consensus output? The second one is the hypothesis alignment accuracy: how does the hypothesis alignment influence the word sequence of the consensus output?

To study the first issue, considering that the word order of CN is decided by the backbone, we performed a set of experiments to compare the influence on consensus output of selecting different backbones for our CN. Table 1 shows the comparison results. We use the WMT09 English-to-French system combination shared task as the evaluation data set, including 2525 sentences and 16 1-best systems. TER is used as the default hypothesis alignment metric. The results are reported in TER, case-sensitive BLEU, NIST (Doddington, 2002) and METEOR (MTR) (Banerjee and Lavie, 2005).

Backbone	**TER**	**BLEU**	**NIST**	**MTR**
Oracle	52.58	33.84	8.04	23.95
Worst Single	69.19	14.73	5.57	12.40
Best Single	59.21	25.43	6.99	18.97
MBR	58.05	26.54	7.12	19.81
Worst-CN	59.16	23.53	7.04	17.63
Best-CN	57.03	26.73	7.29	19.84
MBR-CN	**56.84**	**27.56**	**7.33**	**20.33**

Table 1. The influence of backbone choice on CN

The Worst-CN, Best-CN and MBR-CN are the outputs of the CNs using the worst single, best single and MBR result as the backbone, respectively. We can see that (i) MBR is better than the best single system; and (ii) the MBR-CN obtains the best performance in terms of the four automatic evaluation metrics. The better the word order in the backbone is, the better the translation performance is.

By using the same backbone but different hypothesis alignment methods, we compare the results to address the second issue. Correctly aligning synonyms to each other is a challenging issue. For instance, in Figure 1 (b), "risen" in E_1 and "increased" in E_2 and E_3 express the same meaning with different morphologies. Of course, a simple 'exact match' algorithm is incapable of dealing with this issue. In this experiment, three dominant types of hypothesis alignment metrics are used, namely TER, HMM and IHMM. The data set we used is still the WMT09 English-to-French system combination shared task. TER aligns the words based on the exact match principle; HMM uses the same principle as the word alignment model in (Vogel et al., 1996), while IHMM uses two similarity models and one distortion model to perform the alignment. Table 2 shows the results for these three metrics.

Alignment	**TER**	**BLEU**	**NIST**	**MTR**
TER	56.84	27.56	7.33	20.33
IHMM	56.83	27.27	7.24	20.27
HMM	**56.56**	**27.64**	**7.38**	**20.52**

Table 2. The influence of alignment metrics on CN

In this experiment, the three CNs are built on the MBR-based backbone, and decoded using the same features and weights. We can see that in this task, the HMM approach outperforms the other two methods across all four metrics. When we manually examine the alignment result, the HMM method has a higher word alignment accuracy and produces a lower non-grammatical error rate.

3 Summary of Four Hypothesis Alignment Metrics

Hypothesis alignment is essentially an optimization problem on word alignment. The objective function is to search for the best path of word alignment links between the source sentence F and the target sentence E.

In this section, we will discuss four hypothesis alignment metrics commonly used in our system combination framework.

3.1 TER

The TER (translation error rate) metric measures the ratio of the number of edit operations between the hypothesis $E^{'}$ and the reference E_b to the total number of words in the E_b. Here the backbone E_b is assumed as the reference. The allowable edits include insertions (Ins), deletions (Del), substitutions (Sub) and phrase shifts (Shft). The TER of $E^{'}$ compared to E_b is computed as in Equation (1):

$$TER(E', E_b) = \frac{Ins + Del + Sub + Shft}{N_b} \times 100\% \tag{1}$$

where N_b is the total number of words in E_b. The difference between TER and classical Edit Distance (or WER) (Levenshtein, 1966) is the sequence shift operation, which allows phrasal shifts in the output to be captured.

TER was originally developed as a translation quality evaluation metric, rather than an alignment metric *per se*. Additionally, the working mechanism is also different from the traditional word alignment principle which uses a probabilistic model. However, the editing process still can be regarded as a word alignment process and the objective is to find an optimal or sub-optimal (if the search gets stuck at a local optimum) path. The *Shft* edit is carried out by a greedy algorithm and restricted by three constraints: 1) The shifted words must exactly match the reference words in the destination position. 2) The word sequence of the hypothesis in the original position and the corresponding reference words must not exactly match. 3) The word sequence of the reference that corresponds to the destination position must be misaligned before the shift (Snover et al., 2006).

3.2 HMM

The HMM-based hypothesis alignment model was presented in (Matusov et al., 2006). The idea is to consider alignment between the backbone sentence and the hypothesis sentence as a hidden variable in the conditional probability $P_r(E'|E_b)$. Given the backbone sentence $E_b = \{e_1, \ldots, e_I\}$ and the hypothesis sentence $E' = \{e_1', \ldots, e_J'\}$, which are both the same language, the alignment A between E_b and E' is defined as in Equation (2):

$$P_r(E' \mid E_b) = \sum_A P_r(E', A \mid E_b) \tag{2}$$

where $A \subseteq \{(j,i) : 1 \le j \le J; 1 \le i \le I\}$, i and j represent the word position in E_b and E' respectively. Hence, the alignment issue is to seek the optimum alignment $\hat{A}$ such that:

$$\hat{A} = \arg\max_A P(A \mid e_1^I, e'^J_1) \tag{3}$$

For the HMM-based model, equation (2) can be represented as in (4):

$$P_r(e'^J_1 \mid e_1^I) = \sum_{a_1^J} \prod_1^J [p(a_j \mid a_{j-1}, I) \cdot p(e'_j \mid e_{a_j})] \tag{4}$$

where $p(a_j \mid a_{j-1}, I)$ is the alignment probability and $p(e'_j \mid e_i)$ is the translation probability.

The model parameters are trained iteratively using the GIZA++ toolkit (Och and Ney, 2003) which utilises the maximum likelihood estimation (MLE). Training is performed in the directions $E' \rightarrow E_b$ and $E_b \rightarrow E'$. The final alignment can be determined using cost matrices (Matusov et al., 2006) or by symmetrising, the so-called 'refined' method (Och and Ney, 2003).

3.3 IHMM

The IHMM-based (Indirect Hidden Markov) hypothesis alignment model was proposed in (He et al., 2008) and provides a different way to estimate the synonym matching and word ordering compared to the regular HMM method. In this approach, the parameters of the alignment model are estimated indirectly from a variety of functions, which use an interpolated similarity model p_{sim} to compute the translation probability $p(e'_j \mid e_i)$ and a distance-based distortion model p_d to obtain the alignment probability $p(a_j \mid a_{j-1}, I)$. Therefore, the IHMM model can be written as in (5):

$$P_r(e'^J_1 \mid e^I_1) = \sum_{a^J_1} \prod_1^J [p_d(a_j \mid a_{j-1}, I) \cdot p_{sim}(e'_j \mid e_{a_j})] \quad (5)$$

The similarity model is a linear interpolation model derived based on both semantic similarity p_{sem} and surface similarity p_{sur}, as in (6):

$$p(e'_j \mid e_i) = \alpha \cdot p_{sem}(e'_j \mid e_i) + (1-\alpha) \cdot p_{sur}(e'_j \mid e_i) \quad (6)$$

where p_{sem} is calculated via the bidirectional lexical probabilities between the foreign words and the target words, and p_{sur} is obtained using the longest matched prefix (LMP) algorithm to measure the string similarity. α is the smoothing factor.

The distortion model estimates the first-order dependencies of word ordering, which assumes that the alignment probabilities $p(a_j \mid a_{j-1}, I)$ depend only on the jump distance $(i - i')$ (Vogel et al., 1996), as in (7):

$$p(i \mid i') = \frac{c(d)}{\sum_{l=1}^{I} c(l-i')} = \frac{c(i-i')}{\sum_{l=1}^{I} c(l-i')} \quad (7)$$

where $\{d = i - i' : -4 \le d \le 6\}$ indicates the distortion parameter.

3.4 Source-Side Context Informed Hypothesis Alignment (SSCI)

The SSCI-based hypothesis alignment model was presented in (Du et al., 2009b). The basic idea behind our SSCI method is to employ the source-side word alignment links and source-side phrase span information to heuristically carry out the hypothesis alignment. As to the source–target word alignment task, the aim of hypothesis alignment is to obtain the best word alignment links between the hypothesis and the backbone. Intuitively, this task has been performed in the process of training GIZA++ (Och and Ney, 2003), extracting the phrases and decoding. However, this kind of alignment information is subsequently abandoned during the translation decoding phase. SSCI is intended to keep the source-side word alignment information and utilise it in the hypothesis alignment phase.

There are three steps for SSCI to align the backbone and other hypotheses. Firstly, in the translation decoding stage, the spans of translated source-side phrases are kept as the hidden word alignment information. Secondly, the phrase table is retrieved to acquire the word alignment links between the source and target phrases. Finally, by mapping the word alignment links between the backbone and the hypothesis based on the same span of a source phrase, associated with a normalisation model, hypothesis alignment and CN building can be performed efficiently. This approach does not need any complicated estimation algorithm, nor does it require additional training data or any other resources.

In the mapping step, assuming E_1 is the selected backbone E_b and E' is the hypothesis, $F = \{f_1,...,f_k\}$ is used as a source-side word (or minimum span), $\Lambda_b = \{A_1,...,A_k\}$ as the set of word alignments between F and the counterpart of E_b, and $\Lambda' = \{A'_1,...,A'_k\}$ as the set of word alignments between F and the corresponding part of E'. A_i and A'_i are represented as a set of alignment pairs $\left\langle f_i, \{e_l,...,e_m\} \,||\, (m \geq l \geq 0) \right\rangle$ and $\left\langle f_i, \{e'_p,...,e'_q\} \,||\, (q \geq p \geq 0) \right\rangle$ respectively, which indicates that each source-side word f_i could be aligned to multiple target words or a *null* word. Mapping Λ_b and Λ' to the word alignment between E_b and E' can be achieved as in equations (8) and (9):

$$\Lambda_b \cap \Lambda' = \{A_1 \cap A'_1,...,A_k \cap A'_k\} \qquad (8)$$

$$A_i \cap A'_i = \left\langle \tilde{E}_i, \tilde{E}'_i \right\rangle \qquad (9)$$

where $\tilde{E}_i$ is a set of words in E_b, and $\tilde{E}'_i$ is the set of words in E', both of which are aligned to f_i.

The normalisation model is described as follows: given a backbone e_1^I consisting of I words $e_1,...,e_I$ and a hypothesis e'^J_1 consisting of J words $e'_1,...,e'_J$, $A_{E \to E'}$ denotes the backbone-to-hypothesis word alignment $a_1^I = (a_1,...,a_i,...,a_I)$ between e_1^I and e'^J_1. Since the similarity model primarily normalises the 1-to-N alignments, $A_{E \to E'}$ can be represented as a set of pairs $a'_j = \left\langle E_j, e'_j \right\rangle$ denoting a link between one single hypothesis word e'_j and several backbone words $E_j = \{a_i = j \,||\, i = m,...,n; I \geq n \geq m \geq 1\}$. If the word e'_j is aligned to a *null* word, the set E_j is empty. Given this notation, we modify equation (6) (equation (2) in (He et al., 2008)) as in (10):

$$\begin{aligned} & p(e'_j \,|\, e_i) = \alpha \cdot p_{lex}(e'_j \,|\, e_i) + (1-\alpha) \cdot p_{sim}(e'_j \,|\, e_i) \\ & \hat{a} = \max_{i=m,...,n} \{p(e'_j \,|\, e_i)\} \end{aligned} \qquad (10)$$

where p_{lex} and p_{sim} denote the lexical alignment probability and the similarity between the backbone word e_i and hypothesis word e'_j respectively. α is the interpolation factor, and $\hat{a}$ is the best 1-to-1 link in the set of 1-to-N alignments.

After bidirectional normalisation has been applied, the intersection rule is employed to acquire the 1-to-1, 1-to-*null* and *null*-to-1 links.

4 Modified Consensus Network MBR Decoding

In order to retain the coherent phrases in the original translations (Sim et al., 2007), it is sometimes better to retain sentence-level consensus rather than creating a new word-level consensus which may distort the fluency of the translation. This approach is defined as ConMBR. Firstly, the consensus network decoding is performed to obtain the combination result E_{con}. Then, the hypothesis in the original translations which has the minimum risk loss with respect to Econ is chosen as the consensus output, as in (11):

$$\hat{E}_{conMBR} = \arg\min_{E'} L(E', E_{con}) \cdot P(E \mid F) \tag{11}$$

where $L(E', E_{con})$ is the loss function under a specific evaluation metric. $P(E \mid F)$ is the posterior probability, usually set to a uniform distribution. Alternatively, it can be trained as a system weight via normalisation.

However, given the Oracle scores in Table 1, we believe that some of the original sentences are better than some of the newly generated consensus sentences. Accordingly, we merge the combination results from the different CNs with the original translations and then use the MBR decoder to again search for the best result. We thus define this method as a modified form of ConMBR (mConMBR).

NIST BLEU-4 (Papineni et al., 2002) is used as the loss function in mConMBR, which is computed as in (12):

$$\begin{aligned} L_{BLEU}(E', E) &= 1 - BLEU(E', E) \\ &= 1 - \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n(E', E)\right) \cdot \gamma(E', E) \end{aligned} \tag{12}$$

where $p_n(E', E)$ is the precision of n-grams in the hypothesis E' given the reference E. $\gamma \in [0,1]$ is a brevity penalty.

Therefore, our mConMBR can be rewritten as in (13):

$$\hat{E}_{mconMBR} = \arg\min_{E'} L(E', E) \cdot P(E \mid F) \tag{13}$$

Here we set the posterior probability $P(E \mid F)$ to be a uniform distribution.

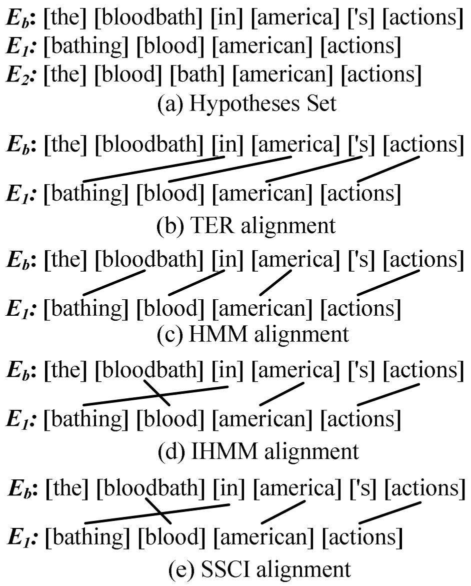
5 Three-pass System Combination Strategy

5.1 Motivation

In recent years, many hypothesis alignment metrics have been proposed using different ways to solve the word alignment issue. The idea of multiple CNs was presented in (Rosti et al., 2007b), where TER is used as the only alignment metric. Considering that the different hypothesis alignment links could bring different combination results, we intend to combine multiple alignment metrics to try to improve translation quality. There are two crucial contributions in our proposed method: (i) we are trying to use the diverse alignment results derived from different hypothesis alignment metrics in a unified combination framework; and (ii) we integrate the super network and mConMBR to combine these alignment metrics and fully make use of the translation results to improve the final quality.

5.2 Description of Algorithm

At sentence level, the different hypothesis alignments could produce different alignment results. As an illustration, in Figure 2(a) E_b is the backbone selected via MBR decoding, and E_1 and E_2 are the original hypotheses from different MT systems. Figures 2(b)–(e) show part of the alignment results performed by TER, HMM, IHMM and SSCI respectively. The alignment links generated by IHMM and SSCI are the same in this example. We can see that the word "*America*" is misaligned to the word "*blood*" by TER in Figure 2(b), while it is correctly aligned to "*American*" by HMM in Figure 2(c), by IHMM in Figure 2(d), and by SSCI in Figure 2(e). It is hard to automatically recognize and evaluate which alignment is better.


E_b: [the] [bloodbath] [in] [america] ['s] [actions]
E_1: [bathing] [blood] [american] [actions]
E_2: [the] [blood] [bath] [american] [actions]
E_b: [the] [bloodbath] [in] [america] ['s] [actions]
E_1: [bathing] [blood] [american] [actions]
E_b: [the] [bloodbath] [in] [america] ['s] [actions]
E_1: [bathing] [blood] [american] [actions]
E_b: [the] [bloodbath] [in] [america] ['s] [actions]
E_1: [bathing] [blood] [american] [actions]
E_b: [the] [bloodbath] [in] [america] ['s] [actions]
E_1: [bathing] [blood] [american] [actions]

(a) Hypotheses Set

(b) TER alignment

(c) HMM alignment

(d) IHMM alignment

(e) SSCI alignment

Figure 2. Hypothesis set and the word alignments performed by different metrics

In order to make full use of the different alignment results and increase the diversity of the search process, we try to combine them in a super network. An example joint network with the priors for each metric and with votes for each arc is shown in Figure 3. According to the word alignment performed by a specific metric, an individual CN can be built with the voting or posterior probability on each arc as shown in Figure 3.

In Figure 3, the super network is constructed by integrating the TER-, HMM-, IHMM- and SSCI-based individual CNs with prior probabilities on the Chinese-to-English translation sentence. At present, the prior probability is manually estimated in light of the performance of each single network. *eps* in Figure 3 is ε that indicates the *null* arc. In our implementation of the four hypothesis alignment methods, SSCI and HMM outperform the other two metrics, and the TER CN is slightly better than the IHMM CN when BLEU score is the MT evaluation function. Accordingly, we set the weights for the four single networks to 0.3, 0.3, 0.25 and 0.15 respectively for the Chinese-to-English task. As for the

English-to-French task, since the hypotheses in the WMT2009 shared task data do not include the source-to-target word alignment information, we just use the TER, HMM and IHMM alignment methods to build a super network, in which the weights are 0.3, 0.5 and 0.2 respectively. Of course, all these weights could be tuned automatically, and we leave for future work an empirical investigation of how these settings might compare to the manually imposed weights. All the individual CNs are connected to a single start node S of ε arcs which contain the prior probabilities. Meanwhile, the CNs are ended by a link of the ε arc to a common end node *E*. The final arcs have a probability of 1.

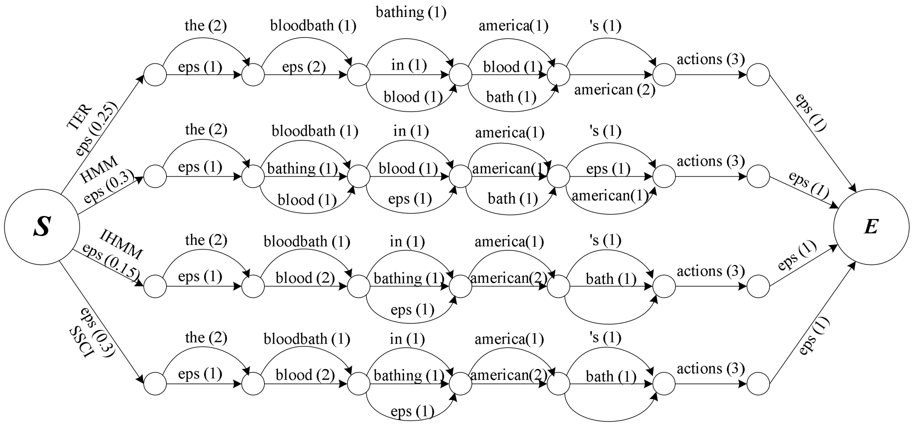


Figure 3. Hypothesis alignment-based multiple confusion networks with prior and posterior probabilities

The construction of the three-pass combination framework may be summarized as follows:

Pass 1: Specific Metric-based Single Network:

1. Merge all the hypotheses from the single MT systems into a new *N*-best list N_s;
2. Utilise the standard MBR decoder to select one hypothesis from the N_s as the backbone;
3. Perform the word alignment between the backbone and the other hypotheses via the TER, HMM, IHMM and SSCI metrics respectively;
4. Carry out word reordering (for all metrics bar TER, which performs reordering in the process of scoring) based on the appropriate word alignments to build CNs of CN_{ter}, CN_{hmm}, CN_{ihmm} and CN_{ssci};
5. Decode the single networks and export the consensus outputs separately.

Pass 2: Super Network:

1) Referring to the 5th step in Pass 1, we train CN_{ter}, CN_{hmm}, CN_{ihmm} and CN_{ssci} via a development set (*devset*) to obtain the weights of each metric-based network, and then estimate the prior probability for each network;
2) Connect the single networks by a start node and an end node to form a multiple hypothesis alignment-based CN;
3) Decode the super network and generate a consensus output.

Pass 3: mConMBR:

1) Combine the N_s with the results from CN_{ter}, CN_{hmm}, CN_{ihmm} and CN_{ssci} as well as the result from the super network to build a new *N*-best list N_{con};
2) Use mConMBR decoding to search for the best final result from N_{con}.

6 Experimental Settings

In this section, we introduce the experimental settings for evaluating and comparing our three-pass alignment-based framework on two different language pairs, namely English-to-French (*E2F*) and Chinese-to-English (*C2E*), in order to measure the effectiveness of our methods. There are two reasons for selecting these two language pairs: (i) one direction is translated out of English and the other direction is translated to English, so we can examine the effects of different directions on system combination strategies; and (ii) English and French are similar languages to some extent, while Chinese and English are very different, both with respect to word order and orthography, so we can investigate the influences of different languages on our proposed combination strategy.

In these two tasks, all the results are reported in terms of BLEU, NIST and METEOR scores. The parameters and weights in the combination process are also optimized with respect to the BLEU score.

6.1 English-to-French Task

In this task, we use the English-to-French data sets in the WMT2009 system combination shared task[1] as the *devset* and the test set. In this direction, the *devset* contains 502 sentences and the test set 2525 sentences. All the sets are from the News domain and have just a single reference translation per source sentence.

There are 16 individual MT systems in this task, each of which have a 1-best and *N*-best list. To save computation effort, we just use the total 1-best results to carry out our experiments.

6.2 Chinese-to-English Task

We trained 5 MT systems to obtain a set of translations. All the MT systems are phrase-based engines, so in order to produce different results with less correlation, we had to train some diverse translation models.

Diversity has a significant influence on the performance of system combination (Macherey and Och, 2007). Although there are many different types of MT systems, if the training data for these systems are the same, there will be a significant correlation between the results. Thus, this would potentially decrease the system combination performance. In order to increase the diversity, we sample the training data to train a number of translation models. Furthermore, we can adjust parameters such as the distortion limit or use different *devsets* to reduce any such correlation.

5 sub-training data sets are randomly sampled from a large database of examples, each of which contains 400K sentence pairs, including the HK, ISI parallel data, UN and other news data.

The *devset* used for translation system parameter training is the NIST MT05 test set which contains 1082 sentences; the *devset* used for system combination parameter tuning (including MBR decoding tuning, CN tuning) is the NIST MT06 test set which contains 1664

[1] http://www.statmt.org/wmt09/system-combination-task.html

sentences. The test set is the NIST MT08 "current" test set which has 1357 sentences from two different domains, namely newswire and web-data genres. All the dev and test sets have 4 reference translations per source sentence.

6.3 System Components

The two language pairs employ the same basic combination framework: the MBR decoder is used to select a potential best hypothesis as the backbone; the CN decoding with 5 features—two language models (one small (from Europarl and Giga data, amounting to about 88 million tokens for English), and one large: 240 million tokens for English, from Europarl, News and News Commentary sources), word posterior probability, *null* word penalty and word penalty—is utilised to build a network and search for the best consensus (Du et al., 2009a). The word alignments between the backbone and the hypothesis are performed by TER, HMM, IHMM and SSCI metrics respectively.

7 Experimental Results and Analysis

7.1 Chinese-to-English Translation

Table 3 first shows the performance of the best and the worst single systems as well as the Oracle result in terms of the BLEU score. In this task, the **SSCI-based** method achieved the best performance in these four individual CNs. The consensus outputs from the **Super_CN** demonstrate a significant improvement by 2.84% BLEU, 4.57% NIST and 0.92% METEOR compared to the **SSCI-based** single network, while the **mConMBR** (which is the final output of the three-pass framework) system did even better, with relative improvements of 4.86% BLEU, 6.32% NIST, and 1.38% METEOR. Moreover, the **mConMBR** also significantly outperforms the **Super_CN**, by 1.97% BLEU, 1.67% NIST, and 0.46% METEOR.

System	BLEU	NIST	MTR
Oracle	26.67	7.93	44.95
Worst Single	17.33	6.59	39.82
Best Single	21.64	6.94	42.95
TER-based	22.47	7.36	43.11
IHMM-based	22.45	7.34	43.20
HMM-based	23.10	7.37	43.27
SSCI-based	**23.25**	**7.44**	**43.35**
Super_CN	**23.91**	**7.78**	**43.75**
mConMBR	**24.38**	**7.91**	**43.95**

Table 3. Results on Chinese-to-English test set

7.2 English-to-French Translation

In Table 4, the **HMM-based** method obtained the best performance among the TER-based, IHMM-based and HMM-based networks on the English-to-French language pair. The consensus outputs from the **Super_CN** obtain a significant improvement by 2.93% BLEU, 4.61% NIST, and 1.61% METEOR compared to the best individual CN system, the **HMM-based** CN. Again, the **mConMBR** does even better, by 4.96% BLEU, 7.59% NIST, 3.46% METEOR, which are statistically significantly better than the results for **Super_CN**, by 1.97% BLEU, 3.66% NIST, and 1.82% METEOR.

System	BLEU	NIST	MTR
Oracle	33.84	8.04	23.95
Worst Single	14.73	5.57	12.40
Best Single	25.43	6.99	18.97
TER-based	27.56	7.33	20.33
IHMM-based	27.27	7.24	20.27
HMM-based	**27.64**	**7.38**	**20.52**
Super_CN	**28.45**	**7.66**	**20.85**
mConMBR	**29.01**	**7.94**	**21.23**

Table 4. Results on English-to-Chinese test set

From these experiments on two language pairs, the results show that (i) both the multiple networks and the mConMBR combination strategy are effective in improving translation quality; and (ii) combining more resources such as in our mConMBR set-up has the capability of improving performance.

7.3 Analysis

From the comparative results conducted on English-to-French and Chinese-to-English, we can see that the multiple-pass combination strategy achieved a significant improvement compared to the individual CN and the best single system.

Different hypothesis alignment metrics can bring different alignment results, which will increase diversity in the search process. Although this might increase the risk of misalignment errors, we can see from the close performance of the multiple individual CNs that this would not impact seriously on translation quality. On the other hand, it can provide more potentially correct candidates for the decoder to determine a final path. By combining different hypothesis alignment results we can construct a multiple word lattice network, which can intrinsically make full use of the context information provided. Regarding the mConMBR method, since the CN is built on the word level, some new sentences can be generated and some new syntactic structures may be brought into the MBR decoding module. For these reasons, the three-pass strategy based on both the super network and the mConMBR are demonstrated to be effective in our experiments.

8 Conclusions and Future Work

In this paper, we investigated four hypothesis alignment metrics used in system combination. Based on these metrics, we presented a unified three-pass framework to combine and utilise the alignment results so as to obtain improved translation performance. We first run the word alignment between the backbone and the hypothesis using the different hypothesis alignment approaches and build the individual CNs according to their respective alignment links, then connect these individual networks with a common start node and a end node to form a super network. Finally, a modified ConMBR is carried out to search for the best final translation from the N_{con} list. Experiments are conducted on Chinese-to-English and English-to-French, and the experimental results clearly demonstrate the effectiveness of our proposed method.

As for future work, firstly we plan to automatically evaluate the alignment quality of different hypothesis alignment metrics. Secondly, we plan to examine how the differences between the hypothesis alignment metrics impact on the accuracy of the super network. We also intend to integrate more alignment metrics into the networks and verify our current findings on other language pairs and translation domains.

9 Acknowledgement

This work is supported by Science Foundation Ireland (http://www.sfi.ie) Grant No. 07/CE/I1142. Thanks to Yanjun Ma for his input on the SSCI approach. Thanks are also due to the reviewers for their insightful comments which served to improve this paper.

10 References

Banerjee, S. and Lavie, A. 2005. *METEOR: an automatic metric for MT evaluation with improved correlation with human judgments*. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, pp. 65-72.

Bangalore, S., Bordel, G., and Riccardi, G. 2001. *Computing consensus translation from multiple machine translation systems*. In Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy, pp. 351-354.

Doddington, G. 2002. *Automatic evaluation of machine translation quality using n-gram cooccurrence statistics*. In Proceedings of Human Language Technology Conference 2002, San Diego, CA, pp. 138-145.

Du, J., He, Y., Penkale, S. and Way, A. 2009a. *MaTrEx: The DCU MT System for WMT2009*. In Proceedings of the Fourth Workshop on Statistical Machine Translation, EACL 2009, Athens, Greece, pp. 95-99.

Du, J., Ma, Y. and Way, A. 2009b. *Source-side context-informed hypothesis alignment for combining outputs from machine translation systems*. In MT Summit XII, Proceedings of the Twelfth Machine Translation Summit, Ottawa, ON, Canada, pp. 230-237.

Du, J. and Way, A. 2009c. *A Three-pass System Combination Framework by Combining Multiple Hypothesis Alignment Methods*. In Proceedings of the International Conference on Asian Language Processing (IALP 2009), Singapore, pp. 172-176.

He, X., Yang, M., Gao, J., Nguyen, P. and Moore, R. 2008. *Indirect HMM-based hypothesis alignment for combining outputs from machine translation systems*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Waikiki, HI, pp. 98-107.

Karakos, D., Eisner, J., Khudanpur, S. and Dreyer, M. 2008. *Machine translation system combination using ITG-based alignments*. In ACL-08: HLT, 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Columbus, OH, pp. 81-84.

Kumar, S. and Byrne, W. 2004. *Minimum Bayes-Risk Decoding for Statistical Machine Translation*. In Proceedings Human-Language Technology and North American Association of Computational Linguistics (HLT-NAACL), Boston, MA, pp. 169-176.

Levenshtein, V. I. 1966. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*. Soviet Physics Doklady 10(8), pp. 707-710.

Macherey, W. and Och, F. 2007. *An empirical study on computing consensus translations from multiple machine translation systems*. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, pp. 986-995.

Matusov, E., Ueffing N., and Ney, H. 2006. *Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment*. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, pp. 33-40.

Och, F. J. and Ney, H. 2002. *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation*. In 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 295-302.

Och, F. J. and Ney, H. 2003. *A systematic comparison of various statistical alignment models*. Computational Linguistics, 29, pp. 19-51.

Papineni, K., Roukos, S., Ward T.and Zhu, W.J. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp. 311-318.

Rosti, A.I., Xiang, B., Matsoukas, S., Schwartz, R., Ayan, N.F., and Dorr, B.J. 2007a. *Combining outputs from multiple machine translation systems*. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, pp. 228-235.

Rosti, A.I., Matsoukas, S. and Schwartz, R. 2007b. *Improved Word-Level System Combination for Machine Translation*. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, pp. 312-319.

Sim, K.C., Byrne, W.J., Gales, M.J.F., Sahbi, H., and Woodland P.C. 2007. *Consensus network decoding for statistical machine translation system combination*. In Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, HI, pp. 105-108.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. 2006. *A study of translation edit rate with targeted human annotation*. In AMTA 2006, Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Visions for the Future of Machine Translation, Cambridge, MA, pp. 223-231.

Vogel, S., Ney, H. and Tillmann, C. 1996. *HMM-based word alignment in statistical translation*. In Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 836-841.