An Augmented Three-Pass System Combination Framework: DCU Combination System for WMT 2010

Jinhua Du, Pavel Pecina, Andy Way

CNGL, School of Computing
Dublin City University
Dublin 9, Ireland
{jdu, ppecina, away}@computing.dcu.ie

Abstract

This paper describes the augmented three-pass system combination framework of the Dublin City University (DCU) MT group for the WMT 2010 system combination task. The basic three-pass framework includes building individual confusion networks (CNs), a super network, and a modified Minimum Bayes-risk (mConMBR) decoder. The augmented parts for WMT2010 tasks include 1) a rescoring component which is used to re-rank the N-best lists generated from the individual CNs and the super network, 2) a new hypothesis alignment metric – TERp – that is used to carry out English-targeted hypothesis alignment, and 3) more different backbone-based CNs which are employed to increase the diversity of the mConMBR decoding phase. We took part in the combination tasks of English-to-Czech and French-to-English. Experimental results show that our proposed combination framework achieved 2.17 absolute points (13.36 relative points) and 1.52 absolute points (5.37 relative points) in terms of BLEU score on English-to-Czech and French-to-English tasks respectively than the best single system. We also achieved better performance on human evaluation.

1 Introduction

In several recent years, system combination has become not only a research focus, but also a popular evaluation task due to its help in improving machine translation quality. Generally, most combination approaches are based on a confusion network (CN) which can effectively re-shuffle the translation hypotheses and generate a new target sentence. A CN is essentially a directed acyclic graph built from 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.

Typically, the dominant CN is constructed at 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 focused on hypothesis alignment due to its significant influence on combination quality.

A multiple CN or "super-network" framework was firstly proposed in Rosti et al. (2007) who used each of all individual system results as the backbone to build CNs based on the same alignment metric, TER (Snover et al., 2006). A consensus network MBR (ConMBR) approach was presented in (Sim et al., 2007), where MBR decoding is employed to select the best hypothesis with the minimum cost from the original single system outputs compared to the consensus output.

Du and Way (2009) proposed a combination strategy that employs MBR, super network, and a modified ConMBR (mConMBR) approach to construct a three-pass system combination framework which can effectively combine different hypothesis alignment results and easily be extended to more alignment metrics. Firstly, a number of individual CNs are built based on different backbones and different kinds of alignment metrics. Each network generates a 1-best output. Secondly, a super network is constructed combining all the individual networks, and a consensus is generated based on a weighted search model. In the third

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 290–295,
Uppsala, Sweden, 15-16 July 2010.

pass, all the 1-best hypotheses coming from single MT systems, individual networks, and the super network are combined to select the final result using the mConMBR decoder.

In the system combination task of WMT 2010, we adopted an augmented framework by extending the strategy in (Du and Way, 2009). In addition to the basic three-pass architecture, we augment our combination system as follows:

- We add a rescoring component in Pass 1 and Pass 2.
- We introduce the TERp (Snover et al., 2009) alignment metric for the English-targeted combination.
- We employ different backbones and hypothesis alignment metrics to increase the diversity of candidates for our mConMBR decoding.

The remainder of this paper is organised as follows. In Section 2, we introduce the three hypothesis alignment methods used in our framework. Section 3 details the steps for building our augmented three-pass combination framework. In Section 4, a rescoring model with rich features is described. Then, Sections 5 and 6 respectively report the experimental settings and experimental results on English-to-Czech and French-to-English combination tasks. Section 7 gives our conclusions.

2 Hypothesis Alignment Methods

Hypothesis alignment plays a vital role in the CN, as the backbone sentence determines the skeleton and the word order of the consensus output.

In the combination evaluation task, we integrated TER (Snover et al., 2006), HMM (Matusov et al., 2006) and TERp (Snover et al., 2009) into our augmented three-pass combination framework. In this section, we briefly describe these three methods.

2.1 TER

The TER (Translation Edit 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 E_b. Here the backbone E_b is assumed to be 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 (1):

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

where N_b is the total number of words in E_b. The difference between TER and Levenshtein edit distance (or WER) is the sequence shift operation allowing phrasal shifts in the output to be captured.

The phrase shift 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).

2.2 HMM

The hypothesis alignment model based on HMM (Hidden Markov Model) considers the alignment between the backbone and the hypothesis as a hidden variable in the conditional probability $P_r(E'|E_b)$. Given the backbone $E_b = \{e_1, \ldots, e_I\}$ and the hypothesis $E' = \{e'_1, \ldots, e'_J\}$, which are both in the same language, the probability $P_r(E'|E_b)$ is defined as in (2):

$$P_r(E'|E_b) = \sum_{A} P_r(E', A|E_b) \quad (2)$$

where the alignemnt $A \subseteq \{(j, i) : 1 \leq j \leq J; 1 \leq i \leq I\}$, 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:

$$\hat{A} = \arg\max_{A} P(A|e_1^I, e_1'^J) \quad (3)$$

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

$$P_r(E'|E_b) = \sum_{a_j^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(e'_j|e_{a_j})] \quad (4)$$

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

2.3 TER-Plus

TER-Plus (TERp) is an extension of TER that aligns words in the hypothesis and reference not only when they are exact matches but also when the words share a stem or are synonyms (Snover et al., 2009). In addition, it uses probabilistic phrasal substitutions to align phrases in the hypothesis and reference. In contrast to the use of

the constant edit cost for all operations such as shifts, insertion, deleting or substituting in TER, all edit costs in TERp are optimized to maximize correlation with human judgments.

TERp uses all the edit operations of TER – matches, insertions, deletions, substitutions, and shifts – as well as three new edit operations: stem matches, synonym matches, and phrase substitutions (Snover et al., 2009). TERp employs the Porter stemming algorithm (Porter, 1980) and WordNet (Fellbaum, 1998) to perform the "stem match" and "synonym match" respectively. Sequences of words in the reference are considered to be paraphrases of a sequence of words in the hypothesis if that phrase pair occurs in the TERp phrase table (Snover et al., 2009).

In our experiments, TERp was used for the French-English system combination task, and we used the default configuration of optimised edit costs.

3 Augmented Three-Pass Combination Framework

The construction of the augmented three-pass combination framework is shown in Figure 1.

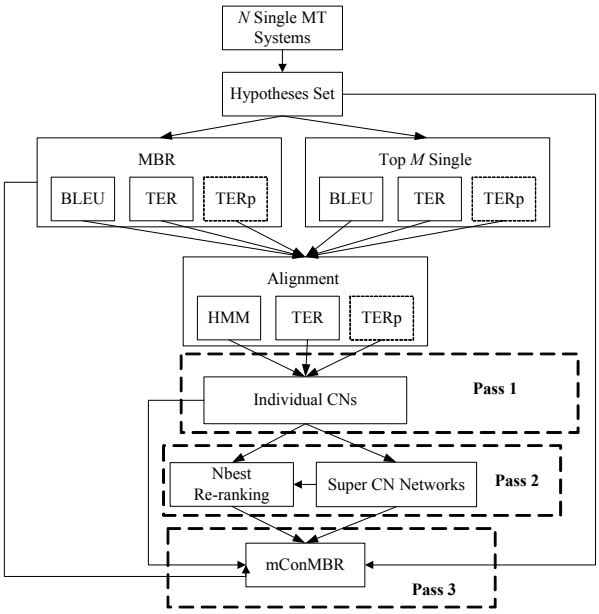

Figure 1: Three-Pass Combination Framework

In Figure 1, the dashed boxes labeled "TERp" indicate that the TERp alignment is only applicable for English-targeted hypothesis alignment. The lines with arrows pointing to "mConMBR" represent adding outputs into the mConMBR decoding component. "Top M Single" indicates that the 1-best results from the best M individual MT systems are also used as backbones to build individual CNs under different alignment metrics. The three dashed boxes represent Pass 1, Pass 2 and Pass 3 respectively. The steps can be summarised as follows:

Pass 1: Specific Metric-based Single Networks

1. Merge all the 1-best hypotheses from single MT systems into a new N-best set N_s.
2. Utilise the standard MBR decoder to select one from the N_s as the backbone given some specific loss function such as TER, BLEU (Papineni et al., 2002) and TERp; Additionally, in order to increase the diversity of candidates used for Pass 2 and Pass 3, we also use the 1-best hypotheses from the top M single MT systems as the backbone. Add the backbones generated by MBR into N_s.
3. Perform the word alignment between the different backbones and the other hypotheses via the TER, HMM, TERp (only for English) metrics.
4. Carry out word reordering based on word alignment (TER and TERp have completed the reordering in the process of scoring) and build individual CNs (Rosti et al., 2007);
5. Decode the single networks and export the 1-best outputs and the N-best lists separately. Add these 1-best outputs into N_s.

Pass 2: Super-Network

1. Connect the single networks using a start node and an end node to form a super-network based on multiple hypothesis alignment and different backbones. In this evaluation, we set uniform weights for these different individual networks when building the super network(Du and Way, 2009).
2. Decode the super network and generate a consensus output as well as the N-best list. Add the 1-best result into N_s.
3. Rescore the N-best lists from all individual networks and super network and add the new 1-best results into N_s.

Pass 3: mConMBR

1. Rename the set N_s as a new set N_{con};
2. Use mConMBR decoding to search for the best final result from N_{con}. In this step, we set a uniform distribution between the candidates in N_{con}.

4 Rescoring Model

We adapted our previous rescoring model (Du et al., 2009) to larger-scale data. The features we used are as follows:

- Direct and inverse IBM model;
- 4-gram and 5-gram target language model;
- 3, 4, and 5-gram Part-of-Speech (POS) language model (Schmid, 1994; Ratnaparkhi, 1996);
- Sentence-length posterior probability (Zens and Ney, 2006);
- N-gram posterior probabilities within the N-best list (Zens and Ney, 2006);
- Minimum Bayes Risk cost. This process is similar to the calculation of the MBR decoding in which we take the current hypothesis in the N-best list as the "backbone", and then calculate and sum up all the Bayes risk cost between the backbone and each of the rest of the N-best list using BLEU metric as the loss function;
- Length ratio between source and target sentence.

The weights are optimized via the MERT algorithm (Och, 2003).

5 Experimental Settings

We participated in the English–Czech and French–English system combination tasks.

In our system combination framework, we use a large-scale monolingual data to train language models and carry out POS-tagging.

5.1 English-Czech

Training Data

The statistics of the data used for language models training are shown in Table 1.

Corpus	*Monolingual tokens (Cz)*	*Number of sentences*
News-Comm	2,214,757	84,706
CzEng	81,161,278	8,027,391
News	205,600,053	13,042,040
Total	288,976,088	21,154,137

Table 1: Statistics of data in the En–Cz task

All the data are provided by the workshop organisers. [1] In Table 1, "News-Comm" indicates the data set of News-Commentary v1.0 and "CzEng" is the Czech–English corpus v0.9 (Bojar and Žabokrtský, 2009). "News" is the Czech monolingual News corpus.

As to our CN and rescoring components, we use "News-Comm+CzEng" to train a 4-gram language model and use "News-Comm+CzEng+News" to train a 5-gram language model. Additionally, we perform POS tagging (Hajič, 2004) for 'News-Comm+CzEng+News" data, and train 3-gram, 4-gram, and 5-gram POS-tag language models.

Devset and Testset

The devset includes 455 sentences and the testset contains 2,034 sentences. Both data sets are provided by the workshop organizers. Each source sentence has only one reference. There are 11 MT systems in the En-Cz track and we use all of them in our combination experiments.

5.2 French-English

Training Data

The statistics of the data used for language models training and POS tagging are shown in Table 2.

Corpus	*Monolingual tokens (En)*	*Number of sentences*
News-Comm	2,973,711	125,879
Europarl	50,738,215	1,843,035
News	1,131,527,255	48,648,160
Total	1,184,234,384	50,617,074

Table 2: Statistics of data in the Fr–En task

"News" is the English monolingual News corpus. We use "News-Comm+Europarl" to train a 4-gram language model and use "News-Comm+Europarl+News" to train a 5-gram language model. We also perform POS tagging (Ratnaparkhi, 1996) for all available data, and train 3-gram, 4-gram and, 5-gram POS-tag language models.

Devset and Testset

We also use all the 1-best results to carry out system combination. There are 14 MT systems in the Fr-En track and we use all of them in our combination experiments.

6 Experimental Results

In this section, all the results are reported on devsets in terms of BLEU and NIST scores.

6.1 English–Czech

In this task, we only used one hypothesis alignment method – TER – to carry out hypothesis

[1] http://www.statmt.org/wmt10/translation-task.html

alignment. However, in order to increase diversity for our 3-pass framework, in addition to using the output from MBR decoding as the backbone, we also separately selected the top 4 individual systems (SYS1, SYS4, SYS6, and SYS11 in our system set) in terms of BLEU scores on the devset as the backbones so that we can build multiple individual CNs for the super network. All the results are shown in Table 3.

SYS	BLEU4	NIST
Worst	9.09	3.83
Best	**17.28**	4.99
SYS1	15.11	4.76
SYS4	12.67	4.40
SYS6	17.28	4.99
SYS11	15.75	4.81
CN-SYS1	17.36	5.12
CN-SYS4	16.94	5.10
CN-SYS6	17.91	5.13
CN-SYS11	17.45	5.09
CN-MBR	**18.29**	5.15
SuperCN	**18.44**	5.17
mConMBR-BAS	**18.60**	5.18
mConMBR-New	**18.84**	5.11

Table 3: Automatic evaluation of the combination results on the En-Cz devset.

"Worst" indicates the 1-best hypothesis from the worst single system, the "Best" is the 1-best hypothesis from the best single system (SYS11)). "CN-SYSX" denotes that we use SYSX ($X = 1, 4, 6, 11$ and MBR) as the backbone to build an individual CN. "mConMBR-BAS" stands for the original three-pass combination framework without rescoring component, while "mConMBR-New" indicates the proposed augmented combination framework. It can be seen from Table 3 that 1) in all individual CNs, the CN-MBR achieved the best performance; 2) SuperCN and mConMBR-New improved by 1.16 (6.71% relative) and 1.56 (9.03% relative) absolute BLEU points compared to the best single MT system. 3) our new three-pass combination framework achieved the improvement of 0.24 absolute (1.29% relative) BLEU points than the original framework.

The final results on the test set are shown in Table 4.

SYS	BLEU4	human eval.(%win)
Best	**16.24**	70.38
mConMBR-BAS	17.91	-
mConMBR-New	**18.41** [2]	**75.17**

Table 4: Evaluation of the combination results on the En-Cz testset.

It can be seen that our "mConMBR-New" framework performs better than the best single system and our original framework "mConMBR-BAS" in terms of automatic BLEU scores and human evaluation for the English-to-Czech task. In this task campaign, we achieved top 1 in terms of the human evaluation.

6.2 French–English

We used three hypothesis alignment methods – TER, TERp and HMM – to carry out word alignment between the backbone and the rest of the hypotheses. Apart from the backbone generated from MBR, we separately select the top 5 individual systems (SYS1, SYS10, SYS11, SYS12, and SYS13 in our system set) respectively as the backbones using HMM, TER and TERp to carry out hypothesis alignment so that we can build more individual CNs for the super network to increase the diversity of candidates for mConMBR. The results are shown in Table 5.[3]

SYS	BLEU4(%)	NIST
Worst	15.04	4.97
Best	**28.88**	6.71
CN-SYS1-TER	29.56	6.78
CN-SYS1-HMM	29.60	6.84
CN-SYS1-TERp	**29.77**	6.83
CN-MBR-TER	30.16	6.91
CN-MBR-HMM	30.19	6.92
CN-MBR-TERp	**30.27**	6.92
SuperCN	**30.58**	6.90
mConMBR-BAS	30.74	7.01
mConMBR-New	**31.02**	6.96

Table 5: Automatic evaluation of the combination results on the Fr-En devset.

"CN-MBR-X" represents the different possible hypothesis alignment methods ($X = \{$TER, HMM, TERp$\}$) which are used to build individual CNs using the output from MBR decoding as the backbone. We can see that the SuperCN and mConMBR-New respectively improved by 1.7 absolute (5.89% relative) and 2.88 absolute (9.97% relative) BLEU points compared to the best single system. Furthermore, our augmented framework "mConMBR-New" achieved the improvement of 0.28 absolute (0.91% relative) BLEU points than the original three-pass framework as well.

[2]This score was measured in-house on the reference provided by the organizer using metric mteval-v13 (ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v13.pl).

[3]In this Table, we take SYS1 as an example to show the results using a single MT system as the backbone under the three alignment metrics.

The final results on the test set are shown in Table 6.

SYS	BLEU4	human eval.(%win)
Best	**28.30**	66.84
mConMBR-BAS	29.21	-
mConMBR-New	**29.82** [2]	**72.15**

Table 6: Evaluation of the combination results on Fr-En test set.

It can be seen that our "mConMBR-New" framework performs the best than the best single system and our original framework "mConMBR-BAS" in terms of automatic BLEU scores and human evaluation for the French–English task.

7 Conclusions and Future Work

We proposed an augmented three-pass multiple system combination framework for the WMT2010 system combination shared task. The augmented parts include 1) a rescoring model to select the potential 1-best result from the individual CNs and super network to increase the diversity for "mConMBR" decoding; 2) a new hypothesis alignment metric "TERp" for English-targeted alignment; 3) 1-best results from the top M individual systems employed to build CNs to augment the "mConMBR" decoding. We took part in the English-to-Czech and French-to-English tasks. Experimental results reported on test set of these two tasks showed that our augmented framework performed better than the best single system in terms of BLEU scores and human evaluation. Furthermore, the proposed augmented framework achieved better results than our basic three-pass combination framework (Du and Way, 2009) as well in terms of automatic evaluation scores. In the released preliminary results, we achieved top 1 and top 3 for the English-to-Czech and French-to-English tasks respectively in terms of human evaluation.

As for future work, firstly we plan to do further experiments using automatic weight-tuning algorithm to tune our framework. 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 to the networks and verify on the other language pairs.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University and has been partially funded by PANACEA, a 7th Framework Research Programme of the European Union (contract number: 7FP-ITC-248064) as well as partially supported by the project *GA405/09/0278* of the Grant Agency of the Czech Republic. Thanks also to the reviewers for their insightful comments.

References

Bojar, O. and Žabokrtský, Z. (2009). CzEng0.9: Large Parallel Treebank with Rich Annotation. *Prague Bulletin of Mathematical Linguistics*, 92.

Du, J., He, Y., Penkale, S., and Way, A. (2009). MaTrEx: The DCU MT System for WMT2009. In *Proceedings of the EACL-WMT 2009*, pages 95–99, Athens, Greece.

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

Fellbaum, C., editor (1998). *WordNet: an electronic lexical database*. MIT Press.

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, volume 1. Charles University Press, Prague.

Kumar, S. and Byrne, W. (2004). Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proceedings of the HLT-NAACL 2004*, pages 169–176, Boston, MA.

Matusov, E., Ueffing, N., and Ney, H. (2006). Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of EACL'06*, pages 33–40.

Och, F. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the ACL-02*, pages 311–318, Philadelphia, PA.

Porter, M. F. (1980). An algorithm for suffix stripping, program.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the EMNLP'96*, pages 133–142, Philadelphia, PA.

Rosti, A., Matsoukas, S., and Schwartz, R. (2007). Improved Word-Level System Combination for Machine Translation. In *Proceedings of ACL'07*, pages 312–319.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Sim, K., Byrne, W., Gales, M., Sahbi, H., and Woodland, P. (2007). Consensus network decoding for statistical machine translation system combination. In *Proceedings of the ICASSP'07*, pages 105–108.

Snover, M., Dorr, B., Schwartz, R., Micciula, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the AMTA'06)*, pages 223–231, Cambridge, MA.

Snover, M., Madnani, N., J.Dorr, B., and Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the WMT'09*, pages 259–268, Athens, Greece.

Zens, R. and Ney, H. (2006). N-gram Posterior Probabilities for Statistical Machine Translation. In *Proceedings of the HLT-NAACL'06)*, pages 72–77, New York, USA.