CONSENSUS NETWORK DECODING FOR STATISTICAL MACHINE TRANSLATION SYSTEM COMBINATION

K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi and P.C. Woodland*

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {kcs23,wjb31,mjfg,hs385,pcw}@eng.cam.ac.uk

ABSTRACT

This paper presents a simple and robust consensus decoding approach for combining multiple Machine Translation (MT) system outputs. A consensus network is constructed from an N-best list by aligning the hypotheses against an alignment reference, where the alignment is based on minimising the translation edit rate (TER). The Minimum Bayes Risk (MBR) decoding technique is investigated for the selection of an appropriate alignment reference. Several alternative decoding strategies proposed to retain coherent phrases in the original translations. Experimental results are presented primarily based on three-way combination of Chinese-English translation outputs, and also presents results for six-way system combination. It is shown that worthwhile improvements in translation performance can be obtained using the methods discussed.

Index Terms— Machine translation, system combination, consensus decoding, Minimum Bayes Risk (MBR) decoding

1. INTRODUCTION

Recently, there have been several successful attempts in combining outputs from multiple Machine Translation (MT) systems. Most of these approaches aim at finding a consensus from a set of alternative translations. For example, Minimum Bayes Risk (MBR) decoding [6] is a hypothesis selection scheme that finds a sentence which yields the lowest *expected loss* (Bayes risk) given an N-best list. This results in a sentence level consensus decoding. In contrast, word-level consensus may be obtained using a consensus network decoding [1, 8], which is similar to techniques in speech recognition designed for hypothesis combination such as ROVER [3] and also to the confusion network decoding method [7]. A consensus network (also known as a "sausage net") comprises a sequence of words, each with alternatives, possibly including nulls, with associated scores. The consensus output is then derived from the network by selecting the word sequence with the best score, where scores can be formed in many different ways such as by voting, or using a posterior probability estimate.

Construction of a word-level consensus network requires the hypotheses in the N-best list to be aligned at the word level. Therefore, the key ingredient in constructing such a network is the alignment process. A simple alignment approach is to select an alignment reference from the N-best list and align the rest of the hypotheses with respect to the chosen reference. Previously investigated alignment methods have included Word Error Rate (WER) alignment [1] and `GIZA++` alignment [8]. In this paper, an alignment scheme based on minimum Translation Error Rate (TER) is presented.

The remaining of this paper is organised as follows: First the two MT evaluation metrics adopted in the paper are discussed. Then the Minimum Bayes Risk (MBR) decoding scheme is described. Sec. 4 introduces the consensus network decoding method, and Sec. 5 describes several alternative ways of deriving consensus from a consensus network. Experimental results on both three-way and six-way system combination are presented in Section 6 using translation from both Chinese and Arabic and text and speech sources.

2. MACHINE TRANSLATION METRICS

MT maps a word sequence F in a source language to a word sequence E in the target language. The translation performance is measured relative to a reference E_r as $L(E, E_r)$. There are several commonly used automatic evaluation metrics, two of which used in this papers are the TER [11] and NIST BLEU-4 scores.

The NIST BLEU-4 score is a variant of BLEU [10], which computes the geometric mean of the precision of n-grams between E and E_r and includes a brevity penalty ($\gamma(E, E_r) \leq 1$) if the hypothesis is shorter than the reference.

$$\texttt{BLEU}(E, E_r) = \exp\left(\sum_{n=1}^{N} \log \frac{p_n(E, E_r)}{N} \times 100\right) \gamma(E, E_r) \quad (1)$$

where $p_n(E, E_r)$ is the precision of n-grams in the hypothesis, E given the reference, E_r. Here $N = 4$ is used.

The TER (translation edit rate) [11] score measures the ratio of the number of string edits between the E and E_r to the total number of words in the reference. The allowable edits include insertions (Ins), deletions (Del), substitutions (Sub) and phrase shifts (Shft)[1].

$$\texttt{TER}(E, E_r) = \frac{\text{Ins} + \text{Del} + \text{Sub} + \text{Shft}}{N} \times 100\% \quad (2)$$

where N is the total number of words in the reference. If phrase shifts are not permitted, the TER metric simplifies to the well-known Word Error Rate (WER) measure.

3. MINIMUM BAYES RISK DECODING

Minimum Bayes Risk (MBR) decoding [6] finds a hypothesis with the lowest Bayes risk with respect to all the translations in an N-best list:

$$E_{\texttt{mbr}} = \arg\min_{E'} \sum_{E} P(E|F) L(E, E') \quad (3)$$

K.C. Sim is now with the Institute for Infocomm Research in Singapore.

This work was in part supported by DARPA under the GALE programme via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors gratefully acknowledge the contributions of colleagues at ISI, BBN and University of Edinburgh in making available translation system output.

[1] Phrase shift is the movement of a contiguous block of words from one location in the hypothesis to another

To appear in ICASSP 2007

where the risk is measured as the expected loss, $L(E, E_r)$, over the posterior probability distribution, $P(E|F)$ of all possible translations. This is often approximated using an N-best list as

$$P(E|F) = \frac{P(E,F)}{\sum_{E'} P(E',F)} \quad (4)$$

where $P(E, F)$, the joint probability of the source and target word sequences, may be derived from the scores generated by the SMT systems. If these scores are unavailable or unreliable, the posterior probability distribution may be assumed to be uniform. This may be the case when the N-best list consists of hypotheses generated by multiple systems where the scores may be incompatible.

The major limitations of MBR decoding are: 1) it is restricted to hypothesis selection; 2) the cost of computing the expected loss increases quadratically with the size of the N-best list. These drawbacks can be overcome using consensus network decoding, which will be described next.

4. CONSENSUS NETWORK DECODING

This section describes a word-level consensus network decoding scheme. Given an N-best list, a consensus network is constructed by aligning all the hypotheses against an alignment reference. This scheme is similar to those proposed in [1, 8]. The fundamental difference lies in the alignment methods used. For example, in [1], a modified WER alignment was used. Recently, Matusov *et al*. introduced an enhanced alignment method using `GIZA++` [8]. In this paper, the use of TER alignment is examined. As described in Section 2, TER measures the minimum number of edits (including phrase shifts) between two sentences. These edits also describe the alignment between the two sentences. This method is very similar to WER alignment, but has the flexibility of word re-ordering. Furthermore, this method does not require a complex alignment model, unlike `GIZA++` alignment.

In this paper, a simple all-against-one alignment approach is adopted. This approach is computationally efficient but has a strong bias towards the chosen alignment reference. Therefore, a carefully chosen alignment reference is important to obtain a good translation performance. A simple choice would be the 1-best hypothesis from the 'best' system. However, knowing the best system requires prior knowledge based on the performance on some development data. This can be avoided by selecting the alignment reference using MBR decoding. To reduce the computational cost, it is possible to perform MBR decoding using a smaller N-best list and then construct the consensus network with a larger list.

After alignment, similar words being aligned together are merged so that a concensus network comprising a sequence of unique word alternatives is formed. Each word is assigned a score based on a simple voting scheme. Empty arcs (ϵ's) are used to accommodate insertions and deletions. Here, a simple example is provided for illustration. Given an N-best list ($N = 4$):

I like eating chocolate ice-cream .
I like to eat vanilla ice-cream .
I like to eat ice-cream with chocolate .
I like ice-cream .

If the first hypothesis was chosen as the alignment reference, the result of alignment may look something like:

I	like	ϵ	eating	chocolate	ϵ	ice-cream	.
I	like	to	eat	vanilla	ϵ	ice-cream	.
I	like	to	eat	chocolate	with	ice-cream	.
I	like	ϵ	ϵ	ϵ	ϵ	ice-cream	.

By merging similar words and performing consensus voting, the resulting consensus network is given by:

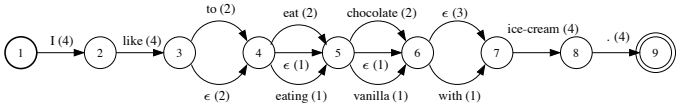

Fig. 1. An example consensus network

Finally, the best path through the network is chosen as the consensus output. In this example, taking the scores as simply the votes for each word alternative, the consensus is given by

`I like to eat chocolate ice-cream.`[2]

Although the alignment and consensus selection process are performed at word-level, the final consensus may exhibit a pseudo phrase-selection characteristic, especially when there are frequently occurring 'pseudo phrases' being aligned consistently. In our example, 'I like' and 'to eat' have been identified as consensual phrases.

5. ALTERNATIVE DECODING STRATEGIES

Word-based consensus decoding may provide a simple solution to deriving 'new' consensual translation not present in the original translations. However, a phrase serves as a more natural translation units semantically. Breaking a coherent phrase has a greater impact on BLEU performance than TER. The form of consensus decoding method detailed in Section 4 tends to insert words into the alignment reference. In practice, the alignment methods are far from ideal, which may lead to an undesirable behaviour of inserting spurious words in between coherent phrases.

To overcome this issue, several alternative deocding strategies were proposed:

Remove insertions (RI): all the words being aligned as insertions are removed before consensus network construction.

Remove phrase breakers (RPB): only remove insertions which are potential phrase breakers. These are identified as insertions in between correctly matched words. In the earlier example, the word 'with' in the third hypothesis may be a phrase breaker since the words that come before and after it have been correctly matched with the alignment reference.

Consensus Network MBR (ConMBR): To retain the coherent phrases in the original translations, it is sometimes better to retain sentence-level consensus rather than creating new word-level consensus which may distort the fluency of the translation. This approach first performs consensus network decoding to obtain, $E_{\texttt{con}}$. The hypothesis in the original translations which has the minimum loss w.r.t. $E_{\texttt{con}}$ is chosen as the consensus output, *i.e.*

$$E_{\texttt{ConMBR}} = \arg\min_{E'} L(E', E_{\texttt{con}}) \quad (5)$$

where $L(E', E_{\texttt{con}})$ is the loss function defined according the appropriate evaluation metric. This is similar to the MBR formulation given in equation (5). Instead of computing the expected loss over a set of possible translations, the Bayes risk is measured with respect to the output from a consensus network decoding. Hence, this method is known as a modified form of MBR decoding.

[2]Note that there is a tie on whether 'to' should be inserted after the second word. The implementation used in this paper favours the one which comes from a hypothesis with a higher rank in the N-best list.

6. EXPERIMENTAL RESULTS

This section presents the experimental results, firstly with three translation systems from ISI. The combination output from this system was evaluated in the context of the NIST `MT06` NIST evaluation [3] as `ISI-CU` system. All of the ISI translation systems were individually tuned to maximise BLEU. In contrast a further set of results were performed on six-way system combination on a set of systems in which some were tuned to maximise BLEU and others TER. Systems were evaluated on the NIST 2004, 2005 and 2006 evaluation test sets with four references for each sentence.

6.1. Three-way Combination

The three way combination was done in the context of "large track" for the NIST `MT06` evaluation. All system combination experiments were conducted based on N-best lists generated by three ISI statistical MT systems: ISI phrase-based system which is similar to [9]; the Hiero system [2]; and a syntax-based system [4]. Unless otherwise stated, performance was measured using TER and BLEU scores based on detokenised lowercase translations.

System	2004		2005	
	TER	BLEU	TER	BLEU
ISI Phrase	55.63	35.89	56.792	33.85
Hiero	**53.64**	39.39	**54.81**	37.06
ISI Syntax	54.20	**39.92**	55.33	**38.17**
MBR-TER	54.23	39.16	55.56	37.09
MBR-BLEU	53.52	39.85	54.30	37.89
+ weights	**53.41**	**40.13**	**54.26**	**38.18**

Table 1. TER/BLEU scores for individual systems and baseline MBR decoding (1-best from each system) on Chinese-English text translation. Weights are tuned on the 2003 test set.

Table 1 shows the TER/BLEU performance of individual systems and MBR decoding of the 1-best translation from each system. `Hiero` and `Syntax` were the best individual systems measured on TER and BLEU metrics respectively. MBR-TER and MBR-BLEU denote MBR decoding using the TER and BLEU loss functions respectively. For MBR decoding, there are only three 1-best hypotheses to select from and the posterior distribution is assumed to be uniform. This was found to yield poorer performance than the best performing individual system. However, it is possible to tune (estimate) the posterior distribution w.r.t. a held-out data (2003 evaluation set). This is equivalent to having system weights that sum to one (2 free parameters). When tuned to maximise the BLEU scores, 0.01–0.21% and 0.82–1.304% absolute improvements in BLEU and TER respectively were obtained.

Alignment Reference	2004		2005	
	TER	BLEU	TER	BLEU
Hiero	52.58	38.52	53.40	37.04
Syntax	52.34	38.97	53.29	37.23
MBR-BLEU (+wgts)	52.30	40.13	53.14	38.53

Table 2. Comparison of alignment references for consensus network decoding using TER alignment for Chinese-English text translation

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

Tables 2, 3 and 4 show the experimental results for consensus network decoding using 10-best hypotheses from each of the three systems. In Table 2, the effect of alignment reference on consensus network decoding was examined. Using the `syntax` system's 1-best hypothesis as alignment reference yields better TER and BLEU performance compared to using the `hiero` 1-best hypothesis. When using the output from MBR-TER and MBR-BLEU (with tuning) as the alignment reference, 0.04–0.14% and 1.16–1.30% improvements on TER and BLEU were obtained over using the `syntax` 1-best output. These results suggest that MBR decoding is useful for alignment reference selection and conveniently eliminates the reliance on the prior knowledge of which is the best performing system.

Alignment Method	2004		2005	
	TER	BLEU	TER	BLEU
WER	52.49	39.33	53.43	37.79
TER	52.30	40.13	53.14	38.53

Table 3. Comparison of alignment methods for consensus network decoding using the MBR-BLEU (+weights) output as alignment reference on Chinese-English text translation

Another important factor which greatly influences the performance of a consensus network decoding scheme is the alignment method used to construct the network. Here, the WER and TER alignment methods were compared in Table 3. It was found that using TER alignment yields better TER (0.2–0.3%) and BLEU (0.7–0.8%) performance than using WER alignment.

Decoding Strategy	2004		2005	
	TER	BLEU	TER	BLEU
Standard	52.30	40.13	53.14	38.53
RI	52.31	39.88	53.09	38.39
RBP	52.33	39.97	53.16	38.51
ConMBR	53.12	40.47	53.91	38.54

Table 4. Consensus network decoding strategies using TER-based alignment on Chinese-English text translation.

Next, various decoding strategies described in Section 5 were compared. The `RI` and `RPB` methods do not seem to help improve the phrase coherency in the original translations, as depicted by the degradation in BLEU scores in Table 4. On the other hand, selecting hypotheses that have the minimum loss w.r.t. the output from a standard consensus network decoding (`ConMBR`) is shown to be beneficial in improving BLEU score performance. However, there is quite a substantial increase in TER (0.77-0.81%). Therefore, while a word-level consensus decoding approach may be suitable for the TER metric, a sentence-level consensus decoding may be better in terms of BLEU scores.

Table 5 shows the performance of `ISI-CU` system combination used in the NIST `MT06` evaluation. Results for both Chinese-English and Arabic-English on two text development sets as well as a broadcast news development set used by the AGILE team in the GALE program[4]. Furthermore the actual evaluation results from the 2006 test are included. The 2006 evaluation data included a mix of newswire, newsgroup and broadcast news (BN) genres while the

[4] The broadcast news development sets used had only one reference translation

Test set	System	Arabic		Chinese	
		TER	BLEU	TER	BLEU
2004	1-best	42.96	46.25	53.64	39.92
	ConMBR	42.06	47.41	53.08	40.51
2005	1-best	38.97	54.57	54.81	38.17
	ConMBR	38.09	55.52	53.84	38.60
BN (speech)	1-best	64.41	21.03	75.40	15.98
	ConMBR	64.13	21.23	75.34	16.34
2006	1-best	51.61	38.81	58.62	32.88
	ConMBR	50.87	39.27	58.12	33.50

Table 5. TER/BLEU scores of MT06 evaluation systems on Arabic-English and Chinese-English text translations. The 2006 are the evaluation results on the entire "NIST" portion of the 2006 test set.

2005 data only includes newswire[5]. The ConMBR method was used as the primary evaluation metric was the BLEU scores. The final N-best list for ConMBR selection was set to be 3×25 for text translation which slowed a slight improvement over just using the 10-best lists from the individual systems. It can be seen that the ConMBR output consistently outperforms the best individual systems based on both the TER and BLEU metrics for all test-sets investigated.

6.2. Six-way Combination

To investigate combination from a larger set of systems, we combined the outputs from six Arabic text-translation systems used with the AGILE team of the DARPA GALE program. In addition to the systems from ISI[6] there were the BBN phrase-based system, the BBN implementation of [2] supplemented with rules containing named entities [12], and the Edinburgh system [5].

The performance on Arabic text translation of the individual systems, MBR-BLEU (unweighted), confusion network decoding and the ConMBR decoding is given in Table 6.2.

System/ Combination	2003		2004	
	TER	BLEU	TER	BLEU
BBN Phrase	41.56	53.32	41.71	45.40
BBN Hiero	42.36	52.03	44.26	42.67
Edinburgh	42.05	52.5	44.20	**47.76**
ISI Hiero	**40.53**	**54.54**	**42.21**	46.49
ISI Phrase	41.94	52.35	43.09	45.21
ISI Syntax	42.96	52.36	45.00	44.11
MBR-BLEU	39.71	56.16	41.29	48.37
Confusion	39.37	55.67	41.21	46.45
ConMBR-BLEU	**39.02**	**56.64**	**40.23**	**48.93**

Table 6. TER and BLEU scores for six individual systems on the 2003 and 2004 Arabic MT NIST test sets, and three alternative combination approaches.

Table 6.2 shows that while there is considerable variability among the performance of the individual systems, the system combination approaches again always yield better performance that the best systems under either the TER or BLEU metrics. On these test sets with the greater diversity in N-Best lists offered by the individual systems, the ConMBR method gives the best performance using both the TER and BLEU measures and is 1.2–2.1 points better than the best individual system for BLEU and 1.5–2.0 points better for TER. Hence, the ConMBR method seems to consistently give good improvements in translation performance.

[5]The 2006 evaluation results are based on mixed case output, the development results are lower-case.

[6]Note that the ISI systems used different training sets and setup to the corresponding systems in 6.1.

7. CONCLUSIONS

This paper has presented a simple and robust system combination method for machine translation by finding consensus from a set of translation hypotheses. This is achieved by constructing a consensus network using TER alignment and a simple voting scheme. Minimum Bayes risk decoding was applied to obtain a single improved reference for efficient alignment. Direct word-level consensus decoding gives promising improvements on TER. However, when eveluated using BLEU, performing hypothesis selection using a Consensus Network MBR decoding scheme yields better results, and this is especially useful when a diverse range of system hypotheses is available for final selection.

8. REFERENCES

[1] S. Bangalore, G. Bordel, & G. Riccardi, "Computing consensus translation from multiple machine translation systems," in *Proc. ASRU*, 2001.

[2] D. Chiang, "A hierarchical phrase-based model for statistical machine translation," in *Proc. ACL*, 2005.

[3] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. IEEE ASRU Workshop*, 1997.

[4] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, & I. Thayer, "Scalable inferences and training of context-rich syntax translation models," in *Proc. ACL*, 2006.

[5] P. Koehn, A. Axelrod, A.B. Mayne, C. Callison-Burch, M. Osborne & David Talbot, "Edinburgh system description for the 2005 IWSLT speech translation evaluation," in *Proc. International Workshop on Spoken Language Translation*, 2005 .

[6] S. Kumar and W. Byrne, "Minimum Bayes-risk decoding for statistical machine translation," in *Proc. of HLT*, 2004.

[7] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1999.

[8] E. Matusov, N. Ueffing, and H. Ney, "Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment," in *Proc. EACL*, 2006.

[9] F.J. Och & H. Ney. "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol.30, no. 2, pp. 417-449, 2004.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," Tech. Rep. RC22176 (W0109-022), IBM Research Division, 2001.

[11] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, & J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. Assoc. for Machine Trans. in the Americas*, 2006.

[12] B. Xiang et al., "The BBN machine translation system for the NIST 2006 MT evaluation," presentation at *NIST Machine Translation Workshop*, 2006.