Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment

Evgeny Matusov, Nicola Ueffing, Hermann Ney
Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University, Aachen, Germany.
{matusov,ueffing,ney}@informatik.rwth-aachen.de

Abstract

This paper describes a novel method for computing a consensus translation from the outputs of multiple machine translation (MT) systems. The outputs are combined and a possibly new translation hypothesis can be generated. Similarly to the well-established ROVER approach of (Fiscus, 1997) for combining speech recognition hypotheses, the consensus translation is computed by voting on a confusion network. To create the confusion network, we produce pairwise word alignments of the original machine translation hypotheses with an enhanced statistical alignment algorithm that explicitly models word reordering. The context of a whole document of translations rather than a single sentence is taken into account to produce the alignment.

The proposed alignment and voting approach was evaluated on several machine translation tasks, including a large vocabulary task. The method was also tested in the framework of multi-source and speech translation. On all tasks and conditions, we achieved significant improvements in translation quality, increasing e. g. the BLEU score by as much as 15% relative.

1 Introduction

In this work we describe a novel technique for computing a consensus translation from the outputs of multiple machine translation systems.

Combining outputs from different systems was shown to be quite successful in automatic speech recognition (ASR). Voting schemes like the ROVER approach of (Fiscus, 1997) use edit distance alignment and time information to create confusion networks from the output of several ASR systems.

Some research on multi-engine machine translation has also been performed in recent years. The most straightforward approaches simply select, for each sentence, one of the provided hypotheses. The selection is made based on the scores of translation, language, and other models (Nomoto, 2004; Paul et al., 2005). Other approaches combine lattices or N-best lists from several different MT systems (Frederking and Nirenburg, 1994). To be successful, such approaches require compatible lattices and comparable scores of the (word) hypotheses in the lattices. However, the scores of most statistical machine translation (SMT) systems are not normalized and therefore not directly comparable. For some other MT systems (e.g. knowledge-based systems), the lattices and/or scores of hypotheses may not be even available.

(Bangalore et al., 2001) used the edit distance alignment extended to multiple sequences to construct a confusion network from several translation hypotheses. This algorithm produces monotone alignments only (i. e. allows insertion, deletion, and substitution of words); it is not able to align translation hypotheses with significantly different word order. (Jayaraman and Lavie, 2005) try to overcome this problem. They introduce a method that allows non-monotone alignments of words in different translation hypotheses for the same sentence. However, this approach uses many heuristics and is based on the alignment that is performed to calculate a specific MT error measure; the performance improvements are reported only in terms of this measure.

Here, we propose an alignment procedure that explicitly models reordering of words in the hypotheses. In contrast to existing approaches, the context of the whole document rather than a single sentence is considered in this iterative, unsupervised procedure, yielding a more reliable alignment.

Based on the alignment, we construct a confusion network from the (possibly reordered) translation hypotheses, similarly to the approach of (Bangalore et al., 2001). Using global system probabilities and other statistical models, the voting procedure selects the best consensus hypothesis from the confusion network. This consensus translation may be different from the original translations.

This paper is organized as follows. In Section 2, we will describe the computation of consensus translations with our approach. In particular, we will present details of the enhanced alignment and reordering procedure. A large set of experimental results on several machine translation tasks is presented in Section 3, which is followed by a summary.

2 Description of the Algorithm

The proposed approach takes advantage of multiple translations for a whole test corpus to compute a consensus translation for each sentence in this corpus. Given a single source sentence in the test corpus, we combine M translation hypotheses $E_1, \ldots, E_M$ from M MT engines. We first choose one of the hypotheses E_m as the primary one. We consider this primary hypothesis to have the "correct" word order. We then align and reorder the other, secondary hypotheses $E_n (n = 1, ..., M; n \neq m)$ to match this word order. Since each hypothesis may have an acceptable word order, we let every hypothesis play the role of the primary translation once, and thus align all pairs of hypotheses $(E_n, E_m);\ n \neq m$.

In the following subsections, we will explain the word alignment procedure, the reordering approach, and the construction of confusion networks.

2.1 Statistical Alignment

The word alignment is performed in analogy to the training procedure in SMT. The difference is that the two sentences that have to be aligned are in the same language. We consider the conditional probability $Pr(E_n|E_m)$ of the event that, given E_m, another hypothesis E_n is generated from the E_m. Then, the alignment between the two hypotheses is introduced as a hidden variable:

$$Pr(E_n|E_m) = \sum_{\mathcal{A}} Pr(E_n, \mathcal{A}|E_m)$$

This probability is then decomposed into the alignment probability $Pr(\mathcal{A}|E_m)$ and the lexicon probability $Pr(E_n|\mathcal{A}, E_m)$:

$$Pr(E_n, \mathcal{A}|E_m) = Pr(\mathcal{A}|E_m) \cdot Pr(E_n|\mathcal{A}, E_m)$$

As in statistical machine translation, we make modelling assumptions. We use the IBM Model 1 (Brown et al., 1993) (uniform distribution) and the Hidden Markov Model (HMM, first-order dependency, (Vogel et al., 1996)) to estimate the alignment model. The lexicon probability of a sentence pair is modelled as a product of single-word based probabilities of the aligned words.

The training corpus for alignment is created from a test corpus of N sentences (usually a few hundred) translated by all of the involved MT engines. However, the effective size of the training corpus is larger than N, since all pairs of different hypotheses have to be aligned. Thus, the effective size of the training corpus is $M \cdot (M-1) \cdot N$. The single-word based lexicon probabilities $p(e_n|e_m)$ are initialized with normalized lexicon counts collected over the sentence pairs (E_n, E_m) on this corpus. Since all of the hypotheses are in the same language, we count co-occurring equal words, i. e. if e_n is the same word as e_m. In addition, we add a fraction of a count for words with identical prefixes. The initialization could be furthermore improved by using word classes, part-of-speech tags, or a list of synonyms.

The model parameters are trained iteratively in an unsupervised manner with the EM algorithm using the GIZA++ toolkit (Och and Ney, 2003). The training is performed in the directions $E_n \rightarrow E_m$ and $E_m \rightarrow E_n$. The updated lexicon tables from the two directions are interpolated after each iteration.

The final alignments are determined using cost matrices defined by the state occupation probabilities of the trained HMM (Matusov et al., 2004). The alignments are used for reordering each secondary translation E_n and for computing the confusion network.

Figure 1: Example of creating a confusion network from monotone one-to-one word alignments (denoted with symbol |). The words of the primary hypothesis are printed in bold. The symbol $ denotes a null alignment or an ε-arc in the corresponding part of the confusion network.

original hypotheses	**1. would you like coffee or tea**								
	2. would you have tea or coffee								
	3. would you like your coffee or								
	4. I have some coffee tea would you like								
alignment and reordering	would|**would** you|**you** have|**like** coffee|**coffee** or|**or** tea|**tea**								
	would|**would** you|**you** like|**like** your|**$** coffee|**coffee** or|**or** $|**tea**								
	I|**$** would\|**would** you\|**you** like\|**like** have\|**$** some|**$** coffee|**coffee** $|**or** tea|**tea**								
confusion network	**$**	**would**	**you**	**like**	**$**	**$**	**coffee**	**or**	**tea**
	$	would	you	have	$	$	coffee	or	tea
	$	would	you	like	your	$	coffee	or	$
	I	would	you	like	have	some	coffee	$	tea

2.2 Word Reordering

The alignment between E_n and the primary hypothesis E_m used for reordering is computed as a function of words in the secondary translation E_n with minimal costs, with an additional constraint that identical words in E_n can not be all aligned to the same word in E_m. This constraint is necessary to avoid that reordered hypotheses with e. g. multiple consecutive articles "the" would be produced if fewer articles were used in the primary hypothesis. The new word order for E_n is obtained through sorting the words in E_n by the indices of the words in E_m to which they are aligned. Two words in E_n which are aligned to the same word in E_m are kept in the original order. After reordering each secondary hypothesis E_n, we determine $M-1$ monotone *one-to-one* alignments between E_m and $E_n, n = 1, \ldots, M; n \neq m$. In case of many-to-one connections of words in E_n to a single word in E_m, we only keep the connection with the lowest alignment costs. The one-to-one alignments are convenient for constructing a confusion network in the next step of the algorithm.

2.3 Building Confusion Networks

Given the $M-1$ monotone one-to-one alignments, the transformation to a confusion network as described by (Bangalore et al., 2001) is straightforward. It is explained by the example in Figure 1. Here, the original 4 hypotheses are shown, followed by the alignment of the reordered secondary hypotheses 2-4 with the primary hypothesis 1. The alignment is shown with the | symbol, and the words of the primary hypothesis are to the right of this symbol. The symbol $ denotes a null alignment or an ε-arc in the corresponding part of the confusion network, which is shown at the bottom of the figure.

Note that the word "have" in translation 2 is aligned to the word "like" in translation 1. This alignment is acceptable considering the two translations alone. However, given the presence of the word "have" in translation 4, this is not the best alignment. Yet the problems of this type can in part be solved by the proposed approach, since every translation once plays the role of the primary translation. For each sentence, we obtain a total of M confusion networks and unite them in a single lattice. The consensus translation can be chosen among different alignment and reordering paths in this lattice.

The "voting" on the union of confusion networks is straightforward and analogous to the ROVER system. We sum up the probabilities of the arcs which are labeled with the same word and have the same start and the same end state. These probabilities are the global probabilities assigned to the different MT systems. They are manually adjusted based on the performance of the involved MT systems on a held-out development set. In general, a better consensus translation can be produced if the words hypothesized by a better-performing system get a higher probability. Additional scores like word confidence measures can be used to score the arcs in the lattice.

2.4 Extracting Consensus Translation

In the final step, the consensus translation is extracted as the best path from the union of confu-

Table 1: Corpus statistics of the test corpora.

	BTEC IWSLT04				BTEC CSTAR03		EPPS TC-STAR	
	Chinese	Japanese	English		Italian	English	Spanish	English
Sentences	500				506		1 073	
Running Words	3 681	4 131	3 092	3 176	2 942	2 889	18 896	18 289
Distinct Words	893	979	1 125	1 134	1 028	942	3 302	3 742

sion networks. Note that the extracted consensus translation can be different from the original M translations. Alternatively, the N-best hypotheses can be extracted for rescoring by additional models. We performed experiments with both approaches.

Since M confusion networks are used, the lattice may contain two best paths with the same probability, the same words, but different word order. We extended the algorithm to favor more well-formed word sequences. We assign a higher probability to each arc of the primary (unreordered) translation in each of the M confusion networks. Experimentally, this extension improved translation fluency on some tasks.

3 Experimental Results

3.1 Corpus Statistics

The alignment and voting algorithm was evaluated on both small and large vocabulary tasks. Initial experiments were performed on the IWSLT 2004 Chinese-English and Japanese-English tasks (Akiba et al., 2004). The data for these tasks come from the Basic Travel Expression corpus (BTEC), consisting of tourism-related sentences. We combined the outputs of several MT systems that had officially been submitted to the IWSLT 2004 evaluation. Each system had used 20K sentence pairs (180K running words) from the BTEC corpus for training.

Experiments with translations of automatically recognized speech were performed on the BTEC Italian-English task (Federico, 2003). Here, the involved MT systems had used about 60K sentence pairs (420K running words) for training.

Finally, we also computed consensus translation from some of the submissions to the TC-STAR 2005 evaluation campaign (TC-STAR, 2005). The TC-STAR participants had submitted translations of manually transcribed speeches from the European Parliament Plenary Sessions (EPPS). In our experiments, we used the translations from Spanish to English. The MT engines for this task had been trained on 1.2M sentence pairs (32M running words).

Table 2: Improved translation results for the consensus translation computed from 5 translation outputs on the Chinese-English IWSLT04 task.

BTEC Chinese-English	WER [%]	PER [%]	BLEU [%]
worst single system '04	58.3	46.6	34.6
best single system* '04	54.6	42.6	40.3
consensus of 5 systems from 2004	47.8	38.0	46.2
system (*) in 2005	50.3	40.5	45.1

Table 1 gives an overview of the test corpora, on which the enhanced hypotheses alignment was computed, and for which the consensus translations were determined. The official IWSLT04 test corpus was used for the IWSLT 04 tasks; the CSTAR03 test corpus was used for the speech translation task. The March 2005 test corpus of the TC-STAR evaluation (verbatim condition) was used for the EPPS task. In Table 1, the number of running words in English is the average number of running words in the hypotheses, from which the consensus translation was computed; the vocabulary of English is the merged vocabulary of these hypotheses. For the BTEC IWSLT04 corpus, the statistics for English is given for the experiments described in Sections 3.3 and 3.5, respectively.

3.2 Evaluation Criteria

Well-established objective evaluation measures like the word error rate (WER), position-independent word error rate (PER), and the BLEU score (Papineni et al., 2002) were used to assess the translation quality. All measures were computed with respect to multiple reference translations. The evaluation (as well as the alignment training) was case-insensitive, without considering the punctuation marks.

3.3 Chinese-English Translation

Different applications of the proposed combination method have been evaluated. First, we focused on combining different MT systems which have the same source and target language. The initial experiments were performed on the BTEC Chinese-English task. We combined translations produced by 5 different MT systems. Table 2 shows the performance of the best and the worst of these systems in terms of the BLEU score. The results for the consensus translation show a dramatic improvement in translation quality. The word error rate is reduced e. g. from 54.6 to 47.8%. The research group which had submitted the best translation in 2004 translated the same test set a year later with an improved system. We compared the consensus translation with this new translation (last line of Table 2). It can be observed that the consensus translation based on the MT systems developed in 2004 is still superior to this 2005 single system translation in terms of all error measures.

We also checked how many sentences in the consensus translation of the test corpus are different from the 5 original translations. 185 out of 500 sentences (37%) had new translations. Computing the error measures on these sentences only, we observed significant improvements in WER and PER and a small improvement in BLEU with respect to the original translations. Thus, the quality of previously unseen consensus translations as generated from the original translations is acceptable.

In this experiment, the global system probabilities for scoring the confusion networks were tuned manually on a development set. The distribution was $0.35, 0.25, 0.2, 0.1, 0.1$, with 0.35 for the words of the best single system and 0.1 for the words of the worst single system. We observed that the consensus translation did not change significantly with small perturbations of these values. However, the relation between the probabilities is very important for good performance. No improvement can be achieved with a uniform probability distribution – it is necessary to penalize translations of low quality.

3.4 Spanish-English Translation

The improvements in translation quality are also significant on the TC-STAR EPPS Spanish-English task. Here, we combined four different systems which performed best in the TC-STAR 2005 evaluation, see Table 3. Compared to the best performing single system, the consensus hypothesis reduces the WER from 41.0 to 39.1%. This result is further improved by rescoring the N-best lists derived from the confusion networks (N=1000). For rescoring, a word penalty feature, the IBM Model 1, and a 4-gram target language model were included. The linear interpolation weights of these models and the score from the confusion network were optimized on a separate development set with respect to word error rate.

Table 3: Improved translation results for the consensus translation computed from 4 translation outputs on the Spanish-English TC-STAR task.

EPPS Spanish-English	WER [%]	PER [%]	BLEU [%]
worst single system	49.1	38.2	39.6
best single system	41.0	30.2	47.7
consensus of 4 systems	39.1	29.1	49.3
+ rescoring	38.8	29.0	50.7

Table 4 gives examples of improved translation quality by using the consensus translation as derived from the rescored N-best lists.

3.5 Multi-source Translation

In the IWSLT 2004 evaluation, the English reference translations for the Chinese-English and Japanese-English test corpora were the same, except for a permutation of the sentences. Thus, we could combine MT systems which have different source and the same target language, performing multi-source machine translation (described e. g. by (Och and Ney, 2001)). We combined two Japanese-English and two Chinese-English systems. The best performing system was a Japanese-English system with a BLEU score of 44.7%, see Table 5. By computing the consensus translation, we improved this score to 49.6%, and also significantly reduced the error rates.

To investigate the potential of the proposed approach, we generated the N-best lists ($N = 1000$) of consensus translations. Then, for each sentence, we selected the hypothesis in the N-best list with the lowest word error rate with respect to the multiple reference translations for the sentence. We then evaluated the quality of these "oracle" translations with all error measures. In a contrastive experiment, for each sentence we simply selected

Table 4: Examples of improved translation quality with the consensus translations on the Spanish-English TC-STAR EPPS task (case-insensitive output).

best system	*I also authorised to committees to certain reports*
consensus	*I also authorised to certain committees to draw up reports*
reference	*I have also authorised certain committees to prepare reports*
best system	*human rights which therefore has fought the european union*
consensus	*human rights which the european union has fought*
reference	*human rights for which the european union has fought so hard*
best system	*we of the following the agenda*
consensus	*moving on to the next point on the agenda*
reference	*we go on to the next point of the agenda*

Table 5: Multi-source translation: improvements in translation quality when computing consensus translation using the output of two Chinese-English and two Japanese-English systems on the IWSLT04 task.

BTEC Chinese-English + Japanese-English	WER [%]	PER [%]	BLEU [%]
worst single system	58.0	41.8	39.5
best single system	51.3	38.6	44.7
consensus of 4 systems	44.9	33.9	49.6

Table 6: Consensus-based combination vs. selection: potential for improvement (multi-source translation, selection/combination of 4 translation outputs).

BTEC Chinese-English + Japanese-English	WER [%]	PER [%]	BLEU [%]
best single system	51.3	38.6	44.7
oracle selection	33.3	29.3	59.2
oracle consensus (1000-best list)	27.0	22.8	64.2

the translation with the lowest WER from the original 4 MT system outputs. Table 6 shows that the potential for improvement is significantly larger for the consensus-based combination of translation outputs than for simple selection of the best translation[1]. In our future work, we plan to improve the scoring of hypotheses in the confusion networks to explore this large potential.

3.6 Speech Translation

Some state-of-the-art speech translation systems can translate either the first best recognition hypotheses or the word lattices of an ASR system. It has been previously shown that word lattice input generally improves translation quality. In practice, however, the translation system may choose, for some sentences, the paths in the lattice with many recognition errors and thus produce inferior translations. These translations can be improved if we compute a consensus translation from the output of at least two different speech translation systems. From each system, we take the translation of the single best ASR output, and the translation of the ASR word lattice.

Two different statistical MT systems capable of translating ASR word lattices have been compared by (Matusov and Ney, 2005). Both systems produced translations of better quality on the BTEC Italian-English speech translation task when using lattices instead of single best ASR output. We obtained the output of each of the two systems under each of these translation scenarios on the CSTAR03 test corpus. The first-best recognition word error rate on this corpus is 22.3%. The objective error measures for the 4 translation hypotheses are given in Table 7. We then computed a consensus translation of the 4 outputs with the proposed method. The better performing word lattice translations were given higher system probabilities. With the consensus hypothesis, the word error rate went down from 29.5 to 28.5%. Thus, the negative effect of recognition errors on the translation quality was further reduced.

4 Conclusions

In this work, we proposed a novel, theoretically well-founded procedure for computing a possibly new consensus translation from the outputs of multiple MT systems. In summary, the main con-

[1] Similar "oracle" results were observed on other tasks.

Table 7: Improvements in translation quality on the BTEC Italian-English task through computing consensus translations from the output of two speech translation systems with different types of source language input.

system	input	WER [%]	PER [%]	BLEU [%]
2	correct text	23.3	19.3	65.6
1	a) single best	32.8	28.6	53.9
	b) lattice	30.7	26.7	55.9
2	c) single best	31.6	27.5	54.7
	d) lattice	29.5	26.1	58.2
consensus a-d		28.5	25.0	58.9

tributions of this work compared to previous approaches are as follows:

- The words of the original translation hypotheses are aligned in order to create a confusion network. The alignment procedure explicitly models word reordering.
- A test corpus of translations generated by each of the systems is used for the unsupervised statistical alignment training. Thus, the decision on how to align two translations of a sentence takes the whole document context into account.
- Large and significant gains in translation quality were obtained on various translation tasks and conditions.
- A significant improvement of translation quality was achieved in a multi-source translation scenario. Here, we combined the output of MT systems which have different source and the same target language.
- The proposed method can be effectively applied in speech translation in order to cope with the negative impact of speech recognition errors on translation accuracy.

An important feature of a real-life application of the proposed alignment technique is that the lexicon and alignment probabilities can be updated with each translated sentence and/or text. Thus, the correspondence between words in different hypotheses and, consequently, the consensus translation can be improved overtime.

5 Acknowledgement

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. This work was also in part funded by the European Union under the integrated project TC-STAR – Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

References

Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii. 2004. *Overview of the IWSLT04 Evaluation Campaign*. Int. Workshop on Spoken Language Translation, pp. 1–12, Kyoto, Japan.

S. Bangalore, G. Bordel, G. Riccardi. 2001. *Computing Consensus Translation from Multiple Machine Translation Systems*. IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Italy.

P. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. *The Mathematics of Statistical Machine Translation*. Computational Linguistics, vol. 19(2):263–311.

J. G. Fiscus. 1997. *A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)*. IEEE Workshop on Automatic Speech Recognition and Understanding.

S. Jayaraman and A. Lavie. 2005. *Multi-Engline Machine Translation Guided by Explicit Word Matching*. 10th Conference of the European Association for Machine Translation, pp. 143-152, Budapest, Hungary.

M. Federico 2003. *Evaluation Frameworks for Speech Translation Technologies*. Proc. of Eurospeech, pp. 377-380, Geneva, Switzerland.

R. Frederking and S. Nirenburg. 1994. *Three Heads are Better Than One*. Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany.

E. Matusov, R. Zens, and H. Ney. 2004. *Symmetric Word Alignments for Statistical Machine Translation*. 20th Int. Conf. on Computational Linguistics, pp. 219–225, Geneva, Switzerland.

E. Matusov and H. Ney. 2005. *Phrase-based Translation of Speech Recognizer Word Lattices Using Loglinear Model Combination*. IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 110-115, San Juan, Puerto-Rico.

T. Nomoto. 2004. *Multi-Engine Machine Translation with Voted Language Model*. 42nd Conference of the Association for Computational Linguistics (ACL), pp. 494-501, Barcelona, Spain.

F. J. Och and H. Ney. 2001. *Statistical Multi-Source Translation*. MT Summit VIII, pp. 253-258, Santiago de Compostela, Spain.

F. J. Och and H. Ney. 2003. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, 29(1):19–51.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Annual Meeting of the ACL, pp. 311–318, Philadelphia, PA, USA.

M. Paul, T. Doi, Y. Hwang, K. Imamura, H. Okuma, and E. Sumita. 2005. *Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation*. International Workshop on Spoken Language Translation, pp. 55-62, Pittsburgh, PA, USA.

TC-STAR Spoken Language Translation Progress Report. 2005. http://www.tc-star.org/documents/deliverable/Deliv_D5_Total_21May05.pdf

S. Vogel, H. Ney, and C. Tillmann. 1996. *HMM-based Word Alignment in Statistical Translation*. 16th Int. Conf. on Computational Linguistics, pp. 836–841, Copenhagen, Denmark.