The RWTH System Combination System for WMT 2010

Gregor Leusch and Hermann Ney
RWTH Aachen University
Aachen, Germany
{leusch,ney}@cs.rwth-aachen.de

Abstract

RWTH participated in the System Combination task of the Fifth Workshop on Statistical Machine Translation (WMT 2010). For 7 of the 8 language pairs, we combine 5 to 13 systems into a single consensus translation, using additional n-best reranking techniques in two of these language pairs. Depending on the language pair, improvements versus the best single system are in the range of $+0.5$ and $+1.7$ on BLEU, and between -0.4 and -2.3 on TER. Novel techniques compared with RWTH's submission to WMT 2009 include the utilization of n-best reranking techniques, a consensus true casing approach, a different tuning algorithm, and the separate selection of input systems for CN construction, primary/skeleton hypotheses, HypLM, and true casing.

1 Introduction

The RWTH approach to MT system combination is a refined version of the ROVER approach in ASR (Fiscus, 1997), with additional steps to cope with reordering between different hypotheses, and to use true casing information from the input hypotheses. The basic concept of the approach has been described by Matusov et al. (2006). Several improvements have been added later (Matusov et al., 2008). This approach includes an enhanced alignment and reordering framework. In contrast to existing approaches (Jayaraman and Lavie, 2005; Rosti et al., 2007), the context of the whole corpus rather than a single sentence is considered in this iterative, unsupervised procedure, yielding a more reliable alignment. Majority voting on the generated lattice is performed using prior weights for each system as well as other statistical models such as a special n-gram language model. In addition to lattice rescoring, n-best list reranking techniques can be applied to n best paths of this lattice. True casing is considered a separate step in RWTH's approach, which also takes the input hypotheses into account.

The pipeline, and consequently the description of the pipeline given in this paper, is based on our pipeline for WMT 2009 (Leusch et al., 2009), with several extensions as described.

2 System Combination Algorithm

In this section we present the details of our system combination method. Figure 1 gives an overview of the system combination architecture described in this section. After preprocessing the MT hypotheses, pairwise alignments between the hypotheses are calculated. The hypotheses are then reordered to match the word order of a selected *primary* or *skeleton* hypothesis. From this, we create a lattice which we then rescore using system prior weights and a language model (LM). The single best path in this CN then constitutes the consensus translation; alternatively the n best paths are generated and reranked using additional statistical models. The consensus translation is then true cased and postprocessed.

2.1 Word Alignment

The proposed alignment approach is a statistical one. It takes advantage of multiple translations for a whole corpus to compute a consensus translation for each sentence in this corpus. It also takes advantage of the fact that the sentences to be aligned are in the same language.

For each of the K source sentences in the test corpus, we select one of its translations $E_n, n = 1, \ldots, M$, as the *primary* hypothesis. Then we align the *secondary* hypotheses $E_m (m = 1, \ldots, M; n \neq m)$ with E_n to match the word order in E_n. Since it is not clear which hypothesis should be primary, i. e. has the "best" word order, we let several or all hypothesis play the role of the primary translation, and align all pairs of hypotheses (E_n, E_m); $n \neq m$. In this paper, we denote the number of possible primary hypotheses by N.

The word alignment is *trained* in analogy to the alignment training procedure in statistical MT. The difference is that the two sentences that have to be aligned are in the same language. We use the IBM Model 1 (Brown et al., 1993) and the Hidden Markov Model (HMM, (Vogel et al., 1996))

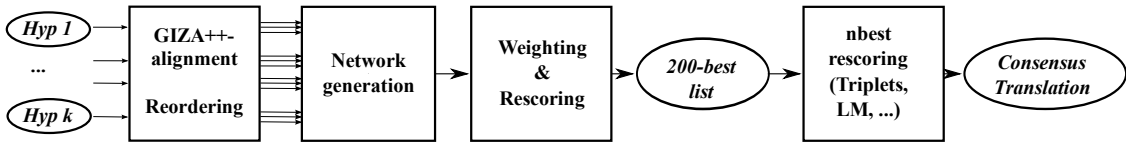

Figure 1: The system combination architecture.

to estimate the alignment model.

The alignment training corpus is created from a test corpus of effectively $N \cdot (M-1) \cdot K$ sentences translated by the involved MT engines. Model parameters are trained iteratively using the GIZA++ toolkit (Och and Ney, 2003). The training is performed in the directions $E_m \rightarrow E_n$ and $E_n \rightarrow E_m$. The final alignments are determined using a cost matrix C for each sentence pair (E_m, E_n). Elements of this matrix are the local costs $C(j,i)$ of aligning a word $e_{m,j}$ from E_m to a word $e_{n,i}$ from E_n. Following Matusov et al. (2004), we compute these local costs by interpolating the negated logarithms of the state occupation probabilities from the "source-to-target" and "target-to-source" training of the HMM model.

2.2 Word Reordering and Confusion Network Generation

After reordering each secondary hypothesis E_m and the rows of the corresponding alignment cost matrix, we determine $M-1$ monotone *one-to-one* alignments between E_n as the primary translation and $E_m, m = 1, \ldots, M; m \neq n$. We then construct the confusion network.

We consider words without a correspondence to the primary translation (and vice versa) to have a null alignment with the empty word ε, which will be transformed to an ε-arc in the corresponding confusion network.

The $M-1$ monotone one-to-one alignments can then be transformed into a confusion network, as described by Matusov et al. (2008).

2.3 Voting in the Confusion Network

Instead of choosing a fixed sentence to define the word order for the consensus translation, we generate confusion networks for N possible hypotheses as primary, and unite them into a single lattice. In our experience, this approach is advantageous in terms of translation quality compared to a minimum Bayes risk primary (Rosti et al., 2007).

Weighted majority voting on a single confusion network is straightforward and analogous to ROVER (Fiscus, 1997). We sum up the probabilities of the arcs which are labeled with the same word and have the same start state and the same end state. This can also be regarded as having a binary system feature in a log-linear model.

2.4 Language Models

The lattice representing a union of several confusion networks can then be directly rescored with an n-gram language model (LM). A transformation of the lattice is required, since LM history has to be memorized.

We train a trigram LM on the outputs of the systems involved in system combination. For LM training, we take the system hypotheses for the same test corpus for which the consensus translations are to be produced. Using this "adapted" LM for lattice rescoring thus gives bonus to n-grams from the original system hypotheses, in most cases from the original phrases. Presumably, many of these phrases have a correct word order. Previous experimental results show that using this LM in rescoring together with a word penalty notably improves translation quality. This even results in better translations than using a "classical" LM trained on a monolingual training corpus. We attribute this to the fact that most of the systems we combine already include such general LMs.

2.5 Extracting Consensus Translations

To generate our consensus translation, we extract the single-best path from the rescored lattice, using "classical" decoding as in MT. Alternatively, we can extract the n best paths for n-best list rescoring.

2.6 n-best-List Reranking

If n-best lists were generated in the previous steps, additional sentence-based features can be calculated on these sentences, and combined in a log-linear way. These scores can then be used to rerank the sentences.

For the WMT 2010 FR–EN and the DE–EN task, we generated 200-best lists, and calculated the following features:

1. Total score from the lattice rescoring
2. NGram posterior weights on those (Zens and Ney, 2006)
3. Word Penalty
4. HypLM trained on a different set of hypotheses (FR–EN only)
5. Large fourgram model trained on Gigaword (DE–EN) or Europarl (FR–EN)
6. IBM1 scores and deletion counts based on a word lexicon trained on WMT training data

7. Discriminative word lexicon score (Mauser et al., 2009)
8. Triplet lexicon score (Hasan et al., 2008)

Other features were also calculated, but did not seem to give an improvement on the DEV set.

2.7 Consensus True Casing

Previous approaches to achieve true cased output in system combination operated on true-cased lattices, used a separate input-independent true caser, or used a general true-cased LM to differentiate between alternative arcs in the lattice, as in (Leusch et al., 2009). For WMT 2010, we use per-sentence information from the input systems to determine the consensus case of each output word. Lattice generation, rescoring, and reranking are performed on lower-cased input, with a lower-cased consensus hypothesis as their result. For each word in this hypothesis, we count how often each casing variant occurs in the input hypotheses for this sentence. We then use the variant with the highest support for the final consensus output. One advantage is that the set of systems used to determine the consensus case does not have to be identical to those used for building the lattice: Assuming that each word from the consensus hypothesis also occurs in one or several of the true casing input hypotheses, we can focus on systems that show a good true casing performance.

3 Tuning

3.1 Tuning Weights for Lattice and n-best Rescoring

For lattice rescoring, we need to tune system weights, LM factor, and word penalty to produce good consensus translations. The same holds for the log-linear weights in n-best reranking.

For the WMT 2010 Workshop, we selected a linear combination of BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) as optimization criterion, $\hat{\Theta} := \text{argmax}_{\Theta} \{BLEU - TER\}$, based on previous experience (Mauser et al., 2008). For more stable results, we use the case-insensitive variants for both measures, despite the explicit use of case information in the pipeline.

System weights were tuned to this criterion using the Downhill Simplex method. Because we considered the number of segments in the tuning set to be too small to allow for a further split into an actual tuning and a control (dev) part, we went for a method closely related to 5-fold cross validation: We randomly split the tuning set into 5 equal-sized parts, and tune parameters on four fifth of the set, measuring progress on the remaining fifth. This was repeated for the other four choices for the "dev" part. Only settings which reliably showed progress on these five different versions were used later on the test set. For the actual weights and numerical parameters to be used on the test set, we calculate the median of the five variants, which lowered the risk of outliers and overfitting.

3.2 System Selection

With the large numbers of input systems – e.g., 17 for DE–EN – and their large spread in translation quality – e.g. 10% abs. in BLEU – not all systems should participate in the system combination process. For the generation of lattices, we considered several variants of systems, often starting from the top, and either replacing some of the systems very similar to others with systems further down the list, or not considering those as primary, adding further systems as additional secondaries.

For true casing, and the additional HypLM for FR–EN, we selected a set of 8 to 12 promising systems, and ran an exhaustive search on all combinations of those to optimize the LM perplexity on the dev set (LM) or the true case BLEU/TER score on a consensus translation (TC). Further research may include a weighted combination here, followed by an optimization of the weights as described in the previous paragraph.

4 Experimental Results

Each language pair and each direction in WMT 2010 had its own set of systems, so we selected and tuned for each direction separately. After submission of our system combination output to WMT 2010, we also calculated scores on the test set (TEST), to validate our results, and as a preparation for this report. Note that the scores reported for DEV are calculated on the full DEV set, but not on any combination of the one-fifth "cross validation" subcorpora.

4.1 FR–EN and EN–FR

For French–English, we selected a set of eight systems for the primary submission, and eleven systems for the contrastive system, of which six served as skeleton. Six different systems were used for an additional HypLM, five for consensus true casing. Table 1 shows the distribution of these systems. We see the results of system combination on DEV and TEST (the latter calculated after submission) in Table 2. System combination itself turns out to have the largest improvement, +0.5 in BLEU and -0.7 in TER on TEST over the best single system. n-best reranking improves this result even more, by +0.3/-0.3. The influence of tuning and of TC selection is measurable on DEV, but rather small on TEST.

For English–French, 13 systems were used to construct the lattice, 5 serving as skeleton. Five different systems were used for true casing. No n-best list reranking was performed here, as preliminary experiments did not show any significant

Table 1: Overview of systems used for FR/EN.

System	FR–EN		EN–FR	
	A	B	A	B
cambridge	P L C	p	P	p
cu-zeman			S	
cmu-statxfer	L	s		
dfki			S	
eu			S	
geneva			S	
huicong		s		
jhu	P L	p	S	p
koc			S	
lig		s		
limsi	P C	p	S C	p
lium	P L C	s	P C	p
nrc	P C	s	S	p
rali	P L	p	P C	p
rwth	P	p	P C	p
uedin	P L C	p	P C	p

"A" is the primary, "B" the contrastive submission.
"P" denotes a system that served as skeleton.
"S" a system that was only aligned to others.
"L" denotes a system used for a larger HypLM-n-best-rescoring.
"C" is a system used for consensus true casing.

Table 2: Results for FR–EN.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
Best single	27.9	55.4	28.5	54.0
Lattice SC	28.4	55.0	29.0	53.3
+ tuning	28.8	54.5	29.1	53.3
+ CV tuning	28.6	54.7	29.1	53.3
+ nbest rerank.	29.0	54.4	29.4	53.0
+ sel. for TC	**29.1**	**54.3**	**29.3**	**53.0**
Contrast. SC	**28.9**	**54.3**	**28.8**	**53.4**

"SC" stands for System Combination output.
"CV" denotes the split into five different tuning and validation parts.
"sel. TC" is the separate selection for consensus true casing.
Systems in bold were submitted for WMT 2010.

Table 3: Results for EN–FR.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
Best single	27.1	55.7	26.5	56.1
Primary SC	**28.3**	**55.2**	**28.2**	**54.7**
Contrast. SC	**28.5**	**54.7**	**28.1**	**54.6**

Table 4: Overview of systems used for DE/EN.

System	DE–EN		EN–DE	
	A	B	A	B
cu-zeman			S	
cmu	C		P	
dfki			S	p
fbk	P C	p	P	
jhu				p
kit	P C	p	P C	p
koc			S C	p
limsi	P	p	P C	p
liu	C		S C	p
rwth	P	p	P C	p
sfu			S	
uedin	P C	p	P C	p
umd	P	p		
uppsala		p	S	

For abbreviations see Table 1.

Table 5: Results for DE–EN.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
Best single	23.8	59.7	23.5	59.7
Lattice SC	24.7	58.5	25.0	57.9
+ tuning	25.1	57.6	25.0	57.6
+ CV tuning	24.8	58.0	24.9	57.8
+ nbest rerank.	25.3	57.6	24.9	57.6
+ sel. for TC	**25.5**	**57.5**	**24.9**	**57.6**
Contrast. SC	**25.2**	**57.7**	**24.8**	**57.7**

For abbreviations see Table 2.

gain in this direction. As a contrastive submission, we submitted the consensus of 8 systems. These are also listed in Table 1. The results can be found in Table 3. Note that the contrastive system was not tuned using the "cross validation" approach; as a result, we expected it to be sensitive to overfitting. We see improvements around +1.7/-1.4 on TEST.

4.2 DE–EN and EN–DE

In the German–English language pair, 17 systems were available, but incorporating only six of them turned out to deliver optimal results on DEV. As shown in Table 4, we used a combination of seven systems in the contrastive submission. While a

Table 6: Results for EN–DE.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
Best single	16.1	66.3	16.4	65.7
Primary SC	**16.4**	**64.9**	**17.0**	**63.7**
Contrast. SC	**16.4**	**64.9**	**17.3**	**63.4**

Table 7: Overview of systems used for CZ/EN.

System	CZ–EN	EN–CZ
aalto	P	
cmu	P C	
cu-bojar	P	P
cu-tecto		S
cu-zeman	P	S C
dcu		P
eurotrans		S
google	P C	P C
koc		P C
pc-trans		S
potsdam		P C
sfu		S
uedin	P C	P C

For abbreviations see Table 1.
No contrastive systems were built for this language pair.

Table 8: Results for CZ–EN and EN–CZ.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
CZ–EN				
Best single	21.8	58.4	22.9	57.5
Primary SC	**22.4**	**59.1**	**23.4**	**57.9**
EN–CZ				
Best single	17.0	67.1	16.6	66.4
Primary SC	**16.7**	**65.4**	**17.4**	**63.6**

different set of five systems was used for consensus true casing, it turned out that using the same six systems for the "additional" HypLM as for the lattice seemed to be optimal in our approach. Table 5 shows the outcome of our experiments: Again, we see that the largest effect on TEST results from system combination as such (+1.5/-1.8). The other steps, in particular tuning and selection for TC, seem to help on DEV, but make hardly a difference on TEST. n-best reranking brings an improvement of -0.2 in TER, but at a minor deterioration (-0.1) in BLEU.

In the opposite direction, English–German, we combined all twelve systems, five of them serving as skeleton. The contrastive submission consists of a combination of eight systems. Six systems were used for true casing. Again, n-best list rescoring did not result in any improvement in preliminary experiments, and was skipped. Results are shown in Table 6: We see that even though both versions perform equally well on DEV (+0.4/-1.4), the contrastive system performs better by +0.3/-0.3 on TEST (+0.9/-2.3).

4.3 CZ–EN and EN–CZ

In both directions involving Czech, the number of systems was rather limited, so no additional selection turned out to be necessary, and we did not build a contrastive system. For Czech–English, all six systems were used; three of them for true casing. For English–Czech, all eleven systems were used in building the lattice, six of them also as skeleton. Five systems were used in the true casing step. Table 7 lists these systems. From the results in Table 8, we see that for CZ–EN, system combination gains around +0.5 in BLEU, but at costs of +0.4 to +0.7 in TER. For EN–CZ, the results look more positive: While we see only -0.3/-1.7 on DEV, there is a significant improvement of +1.2/-2.8 on TEST.

Table 9: Overview of systems used for ES/EN.

System	EN–ES	
	A	B
cambridge	P C	p
dcu	P	p
dfki	P C	p
jhu	P C	p
sfu	P C	p
uedin	P C	p
upv		p
upv-nnlm	P	p

Table 10: Results for EN–ES.

	TUNE		TEST	
	BLEU	TER	BLEU	TER
ES–EN				
Best single	28.7	53.6	–	–
SC	29.0	53.3	–	–
EN–ES				
Best single	27.8	55.2	28.7	54.0
Primary SC	**29.5**	**52.9**	**30.0**	**51.4**
Contrast. SC	**29.6**	**52.8**	**30.1**	**51.7**

4.4 ES–EN and EN–ES

In the Spanish–English language pair, we did not see any improvement at all on the direction with English as target in preliminary experiments. Consequently, and given the time constraints, we did not further investigate on this language pair. Post-eval experiments revealed that improvements of +0.3/-0.3 are possible, with far off-center weights favoring the top three systems.

On English–Spanish, where these preliminary experiments showed a gain, we used seven out of the available ten systems in building the lattice for the primary system, eight for the contrastive. Five of those were uses for consensus true casing. Table 9 lists these systems. Table 10 shows the results on this language pair: For both the primary and the contrastive systems we see improve-

ments of around +1.7/-2.3 on DEV, and +1.3/-2.6 on TEST. Except for the TER on TEST, these two submissions differ only by ± 0.1 from each other.

5 Conclusions

We have shown that our system combination system can lead to significant improvements over single best MT output where a significant number of comparably good translations is available on a single language pair. n-best reranking can further improve the quality of the consensus translation; results vary though. While consensus true casing turned out to be very useful despite of its simplicity, we were unable to find significant improvements on TEST from the selection of a separate set of true casing input systems.

Acknowledgments

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023.

References

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

J. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*.

S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer. 2008. Triplet lexicon models for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Honolulu, Hawaii, October. Association for Computational Linguistics.

S. Jayaraman and A. Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT)*, pages 143–152, Budapest, Hungary, May.

G. Leusch, E. Matusov, and H. Ney. 2009. The RWTH system combination system for WMT 2009. In *Fourth Workshop on Statistical Machine Translation*, pages 56–60, Athens, Greece, March. Association for Computational Linguistics.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric word alignments for statistical machine translation. In *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pages 219–225, Geneva, Switzerland, August.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 33–40, Trento, Italy, April.

E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y. S. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney. 2008. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

A. Mauser, S. Hasan, and H. Ney. 2008. Automatic evaluation measures for statistical machine translation system optimization. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

A. Mauser, S. Hasan, and H. Ney. 2009. Extending statistical machine translation with discriminative and trigger-based lexicon models. In *Conference on Empirical Methods in Natural Language Processing*, pages 210–217, Singapore, August.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

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

A. V. Rosti, S. Matsoukas, and R. Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 312–319, Prague, Czech Republic, June.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proc. of the 7th Conf. of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Boston, MA, August.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

R. Zens and H. Ney. 2006. N-gram posterior probabilities for statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 72–77, New York City, June.