CMU System Combination via Hypothesis Selection for WMT'10

Almut Silja Hildebrand
Carnegie Mellon University
Pittsburgh, USA
silja@cs.cmu.edu

Stephan Vogel
Carnegie Mellon University
Pittsburgh, USA
vogel@cs.cmu.edu

Abstract

This paper describes the CMU entry for the system combination shared task at WMT'10. Our combination method is hypothesis selection, which uses information from n-best lists from the input MT systems, where available. The sentence level features used are independent from the MT systems involved. Compared to the baseline we added source-to-target word alignment based features and trained system weights to our feature set. We combined MT systems for French - English and German - English using provided data only.

1 Introduction

For the combination of machine translation systems there have been several approaches described in recent publications. One uses confusion networks formed along a skeleton sentence to combine translation systems as described in (Rosti et al., 2008) and (Karakos et al., 2008). A different approach described in (Heafield et al., 2009) is not keeping the skeleton fixed when aligning the systems. Another approach selects whole hypotheses from a combined n-best list (Hildebrand and Vogel, 2008).

Our setup follows the latter approach. We combine the output from the submitted translation systems, including n-best lists where available, into one joint n-best list, then calculate a set of features consistently for all hypotheses. We use MER training on the provided development data to determine feature weights and re-rank the joint n-best list. We train to maximize BLEU.

2 Features

For our entries to the WMT'09 we used the following feature groups (in parenthesis are the number of separate feature values per group):

- Language model scores (3)
- Word lexicon scores (6)
- Sentence length features (3)
- Rank feature (1)
- Normalized n-gram agreement (6)
- Source-target word alignment features (6)
- Trained system weights (no. of systems)

The details on language model and word lexicon scores can be found in (Hildebrand and Vogel, 2008) and details on the rank feature and the normalized n-gram agreement can be found in (Hildebrand and Vogel, 2009). We use three sentence length features, which are the ratio of the hypothesis length to the length of the source sentence, the diversion of this ratio from the overall length ratio of the bilingual training data and the difference between the hypothesis length and the average length of the hypotheses in the n-best list for the respective source sentence. The system weights are trained together with the other feature weights during MERT using a binary feature per system. To the feature vector for each hypothesis one feature per input system is added; for each hypothesis one of the features is one, indicating which system it came from, all others are zero.

2.1 Source-Target Word Alignment Features

We trained the IBM word alignment models up to model 4 using the GIZA++ toolkit (Och and Ney, 2003) on the bilingual training corpus. Then a forced alignment algorithm utilizes the trained models to align each source sentence to each translation hypothesis in its respective n-best list.

We use the alignment score given by the word alignment models, the number of unaligned words

Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR, pages 307–310,
Uppsala, Sweden, 15-16 July 2010.

and the number of NULL aligned words, all normalized by the sentence length, as three separate features. We calculate these alignability features for both language directions.

3 Experiments

In the WMT shared translation task only a very small number of participants submitted n-best lists, e.g. in the German-English track there were only four n-best lists among the 16 submissions. Our combination method is proven to work significantly better when n-best lists are available.

For all our experiments on the data from WMT'09, which was available for system combination development as well as the WMT'10 shared task data we used the same setup and the same statistical models.

To train our language models and word lexica we only used provided data. We trained the statistical word lexica on the parallel data provided for each language pair[1]. For each combination we used three language models: a 4-gram language model trained on the English part of the parallel training data, a 1.2 giga-word 3-gram language model trained on the provided monolingual English data, and an interpolated 5-gram language model trained on the English GigaWord corpus. We used the SRILM toolkit (Stolcke, 2002) for training. We chose to train three separate LMs for the three corpora, so the feature weight training can automatically determine the importance of each corpus for this task. The reason for training only a 3-gram LM from the wmt10 monolingual data was simply that there were not sufficient time and resources available to train a bigger model.

For each of the two language pairs we compared a combination that used the word alignment features, or trained system weights or both of these feature groups in addition to the features described in (Hildebrand and Vogel, 2009) which serves a baseline for this set of experiments.

For combination we tokenized and lowercased all data, because the n-best lists were submitted in various formats. Therefore we report the case insensitive scores here. The combination was optimized toward the BLEU metric, therefore TER results might not be very meaningful here and are only reported for completeness.

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

3.1 French-English data from WMT'09

We used 14 systems from the restricted data track of the WMT'09 including five n-best lists. The scores of the individual systems for the combination tuning set range from BLEU 27.93 for the best to 15.09 for the lowest ranked individual system (case insensitive evaluation).

system	tune	test
best single	27.93 / 56.53	27.21 / 56.99
baseline	30.17 / 54.76	28.89 / 55.74
+ wrd al	30.67 / 54.34	28.69 / 55.67
+ sys weights	29.71 / 55.45	28.07 / 56.18
all features	30.30 / 54.53	28.37 / 55.77

Table 1: French-English Results: BLEU / TER

The combination outperforms the best single system by 1.7 BLEU points. Here adding the 14 binary features for training system weights with MERT hurts the combinations performance on the unseen data. The reason for this might be the rather small tuning set of 502 sentences with one reference. Adding the word alignment features does not improve the result either, the difference to the baseline is at the noise level.

3.2 German-English data from WMT'09

For our experiments on the development data for German-English we used the top 12 systems, scoring between BLEU 23.01 and BLEU 16.06, excluding systems known to use data beyond the provided data. Within those 12 system outputs were four n-best lists, three of which were 100-best and one was 10-best.

system	tune	test
best single	23.01 / 60.52	21.44 / 62.33
baseline	26.28 / 58.69	23.62 / 60.49
+ wrd al	26.25 / 59.13	23.42 / 61.11
+ sys weights	26.78 / 58.48	23.28 / 60.80
all features	26.81 / 58.12	23.51 / 60.25

Table 2: German-English Results: BLEU / TER

Our system combination via hypothesis selection could improve translation quality by +2.2 BLEU over the best single system on the unseen test set. Again, the differences between the four different feature sets are not significant on the unseen test set.

3.3 French-English WMT'10 system combination shared task

Out of 14 systems submitted to the French-English translation task, we combined the top 11 systems, the best of which scored 28.58 BLEU and the last 24.16 BLEU on the tuning set. There were only three n-best lists among the submissions. We included up to 100 hypotheses per system in our joint n-best list.

system	tune	test
best sys.	28.58 / 54.17	29.98 / 52.62 / 53.88
baseline	30.67 / 52.62	29.94 / 52.53 / -
+ w. al	30.69 / 52.76	29.97 / 52.76 / 53.76
+ sys w.	30.90 / 52.44	29.79 / 52.84 / 54.05
all feat.	31.10 / 52.06	29.80 / 52.86 / 53.67

Table 3: French-English Results: BLEU / TER / MaxSim

Our system combination via hypothesis selection could not improve the translation quality compared to the best single system on the unseen data. Adding any of the new feature groups to the baseline does not change the result of the combination significantly. This result could be explained by the fact, that due to computational problems and time constraints we were not able to train our models on the whole provided French-English training data. This should only affect the lexicon and word alignment feature groups though.

3.4 German-English WMT'10 system combination shared task

For the German-English combination we used 13 out of the 16 submitted systems, which scored between BLEU 25.01 to BLEU 19.76 on the tuning set. Our combination could improve translation quality by +1.64 BLEU compared to the best system.

system	tune	test
best sys.	25.01 / 58.34	23.89 / 59.14 / 51.10
baseline	26.47 / 56.89	25.44 / 57.96 / -
+ w. al	26.37 / 57.02	25.25 / 58.34 / 50.72
+ sys w.	27.67 / 56.05	25.53 / 57.70 / 51.06
all feat.	27.66 / 56.35	25.25 / 57.86 / 50.83

Table 4: German-English Results: BLEU / TER / MaxSim

The word alignment features seem to hurt performance slightly, which might be due to the more difficult word alignment between German and English compared to other language pairs. But this is not really a strong conclusion, because all differences of the results on the unseen data are not significant.

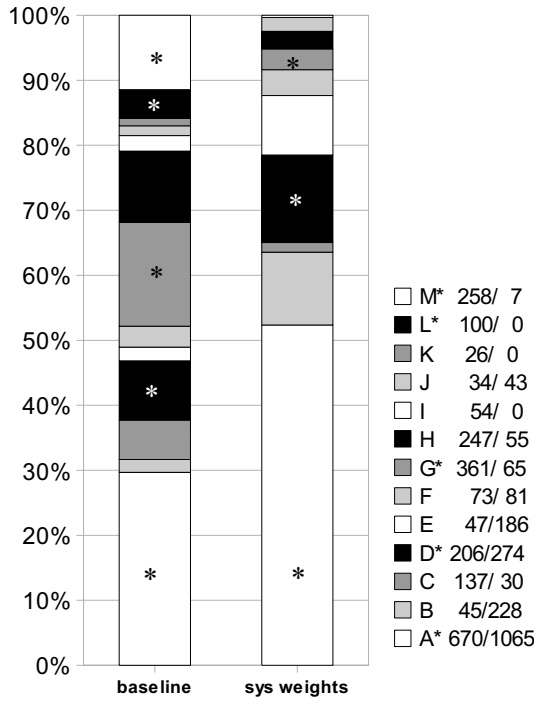

Figure 1: German-English '10: Contributions of the individual systems to the final translation, percentages and absolute number of hyps chosen.

Figure 1 shows, how many hypotheses were contributed by the individual systems to the final translation (unseen data) in the baseline combination compared with the one with trained system weights. The systems A to M are ordered by their BLEU score on the development set. The bars show percentages of the test set, the numbers listed next to the systems A to M give the absolute number of hypotheses chosen from the system for the two depicted combinations. The systems which provided n-best lists, marked with a star in the diagram, clearly dominate the selection in the baseline, but this effect is gone when system weights are used. The dominance of system A in the latter is to be expected, because it is a whole BLEU point ahead of the next ranking system on the system combination tuning set. In the baseline combination identical hypotheses contributed by different systems have an identical total score. In

that case the hypothesis is attributed to all systems which contributed it. This accounts for the higher total number of hypotheses shown in the graphic for the baseline as well as for part of the contributions of the low ranking systems. For example 35 hypotheses were provided identically from two systems and still four hypotheses were produced by all 13 systems, for example the sentence: "aber es geht auch um wirtschaftliche beziehungen ." - "but it is also about economic relations .".

4 Conclusions

In this paper we explored new features in our system combination system, which performs hypothesis selection. We used hypothesis to source sentence alignment scores as well system weight features.

Most systems available for combination did not submit n-best lists, which decreases the effectiveness of our combination method significantly.

The reason for not getting an improvement from word alignment features might be that the top systems might be using more clever word alignment strategies than running the GIZA++ toolkit out of the box. Therefore the alignability according to these weaker models does not give useful ranking information for rescoring.

Experiments on different language pairs and data sets have shown improvements for training system weights in the past for certain setups. Combining up to 14 individual translation systems adds that many features to the feature set for which weights have to optimized via MERT. The provided tuning set of 455 sentences with only one reference is extremely small. It is possible, that MERT could not reliably determine feature weights here. In the setup where this feature set was used successfully, a tuning set of close to 2000 lines with four references was available. It is not possible to improve the tuning data situation by using the provided data from last years workshop as additional tuning data, because the set of systems submitted is not the same and even the systems submitted by the same sites might have changed significantly.

Interesting to note is that looking at the numbers, the German-English combination with an improvement of +1.64 BLEU over the best single system seems to have worked much better than the French-English one with no improvement. But looking at the preliminary human evaluation result the picture is opposite: For German-English our combination is ranked below several of the single systems and most of the combinations, while for French-English it tops the list of all systems and combinations in the workshop.

Acknowledgments

We would like to thank the participants in the WMT'10 shared translation task for providing their data, especially n-best lists. This work was partly funded by DARPA under the project GALE (Grant number #HR0011-06-2-0001).

References

Kenneth Heafield, Greg Hanneman, and Alon Lavie. 2009. Machine translation system combination with flexible word ordering. In *StatMT '09: Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 56–60, Morristown, NJ, USA. Association for Computational Linguistics.

Almut Silja Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *MT at work: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 254–261, Waikiki, Hawaii, October. Association for Machine Translation in the Americas.

Almut Silja Hildebrand and Stephan Vogel. 2009. CMU system combination for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 47–50, Athens, Greece, March. Association for Computational Linguistics.

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, and Markus Dreyer. 2008. Machine translation system combination using itg-based alignments. In *Proceedings of ACL-08: HLT, Short Papers*, pages 81–84, Columbus, Ohio, June. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Antti-Veikko Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 183–186, Columbus, Ohio, June. Association for Computational Linguistics.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings International Conference for Spoken Language Processing*, Denver, Colorado, September.