CMU System Combination for WMT'09

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'09. Our combination method is hypothesis selection, which uses information from n-best lists from several MT systems. The sentence level features are independent from the MT systems involved. To compensate for various n-best list sizes in the workshop shared task including first-best-only entries, we normalize one of our high-impact features for varying sub-list size. We combined restricted data track entries in French - English, German - English and Hungarian - English using provided data only.

1 Introduction

For the combination of machine translation systems there have been two main approaches described in recent publications. One uses confusion network decoding to combine translation systems as described in (Rosti et al., 2008) and (Karakos et al., 2008). The other approach selects whole hypotheses from a combined n-best list (Hildebrand and Vogel, 2008).

Our setup follows the approach described in (Hildebrand and Vogel, 2008). We combine the output from the available translation systems into one joint n-best list, then calculate a set of features consistently for all hypotheses. We use MER training on a development set to determine feature weights and re-rank the joint n-best list.

2 Features

For our entries to the WMT'09 we used the following feature groups:

- Language model score
- Word lexicon scores
- Sentence length features
- Rank feature
- Normalized n-gram agreement

The details on language model and word lexicon scores can be found in (Hildebrand and Vogel, 2008). We use two sentence length features, which are the ratio of the hypothesis length to the length of the source sentence and the difference between the hypothesis length and the average length of the hypotheses in the n-best list for the respective source sentence. We also use the rank of the hypothesis in the original system's n-best list as a feature.

2.1 Normalized N-gram Agreement

The participants of the WMT'09 shared translation task provided output from their translation systems in various sizes. Most submission were 1st-best translation only, some submitted 10-best up to 300-best lists.

In preliminary experiments we saw that adding a high scoring 1st-best translation to a joint n-best list composed of several larger n-best lists does not yield the desired improvement. This might be due to the fact, that hypotheses within an n-best list originating from one single system (sub-list) tend to be much more similar to each other than to hypotheses from another system. This leads to hypotheses from larger sub-lists scoring higher in the n-best list based features, e.g. because they collect more n-gram matches within their sub-list, which "supports" them the more the larger it is.

Previous experiments on Chinese-English showed, that the two feature groups with the highest impact on the combination result are the language model and the n-best list based n-gram agreement. Therefore we decided to focus on the n-best list n-gram agreement for exploring sub-list

Proceedings of the Fourth Workshop on Statistical Machine Translation , pages 47–50,
Athens, Greece, 30 March – 31 March 2009.

size normalization to adapt to the data situation with various n-best list sizes.

The n-gram agreement score of each n-gram in the target sentence is the relative frequency of target sentences in the n-best list for one source sentence that contain the n-gram e, independent of the position of the n-gram in the sentence. This feature represents the percentage of the translation hypotheses, which contain the respective n-gram. If a hypothesis contains an n-gram more than once, it is only counted once, hence the maximum for the agreement score $a(e)$ is 1.0 (100%). The agreement score $a(e)$ for each n-gram e is:

$$a(e) = \frac{C}{L} \tag{1}$$

where C is the count of the hypotheses containing the n-gram and L is the size of the n-best list for this source sentence.

To compensate for the various n-best list sizes provided to us we modified the n-best list n-gram agreement by normalizing the count of hypotheses that contain the n-gram by the size of the sub-list it came from. It can be viewed as either collecting fractional counts for each n-gram match, or as calculating the n-gram agreement percentage for each sub-list and then interpolating them. The normalized n-gram agreement score $a_{norm}(e)$ for each n-gram e is:

$$a_{norm}(e) = \frac{1}{P} \sum_{j=1}^{P} \frac{C_j}{L_j} \tag{2}$$

where P is the number of systems, C_j is the count of the hypotheses containing the n-gram e in the sublist p_j and L_j is the size of the sublist p_j.

For the extreme case of a sub-list size of one the fact of finding an n-gram in that hypothesis or not has a rather strong impact on the normalized agreement score. Therefore we introduce a smoothing factor λ in a way that it has an increasing influence the smaller the sub-list is:

$$a_{smooth}(e) = \frac{1}{P} \sum_{j=1}^{P} \left[\frac{C_j}{L_j} (1 - \frac{\lambda}{L_j}) \right] + \left[\frac{L_j - C_j}{L_j} \frac{\lambda}{L_j} \right] \tag{3}$$

where P is the number of systems, C_j is the count of the hypotheses containing the n-gram in the sublist p_j and L_j is the size of the sublist p_j. We used an initial value of $\lambda = 0.1$ for our experiments.

In all three cases the score for the whole hypothesis is the sum over the word scores normalized by the sentence length. We use n-gram lengths $n = 1..6$ as six separate features.

3 Preliminary Experiments Arabic-English

For the development of the modification on the n-best list n-gram agreement feature we used n-best lists from three large scale Arabic to English translation systems. We evaluate using the case insensitive BLEU score for the MT08 test set with four references, which was unseen data for the individual systems as well as the system combination. Table 1 shows the initial scores of the three input systems.

system	MT08
A	47.47
B	46.33
C	44.42

Table 1: Arabic-English Baselines: BLEU

To compare the behavior of the combination result for different n-best list sizes we combined the 100-best lists from systems A and C and then added three n-best list sizes from the middle system B into the combination: 1-best, 10-best and full 100-best. For each of these four combination options we ran the hypothesis selection using the plain version of the n-gram agreement feature a as well as the normalized version without a_{norm} and with smoothing a_{smooth} .

combination	a	a_{norm}	a_{smooth}
A & C	48.04	48.09	48.13
A & C & B_1	47.84	48.34	48.21
A & C & B_{10}	48.29	48.33	48.47
A & C & B_{100}	48.91	48.95	49.02

Table 2: Combination results: BLEU on MT08

The modified feature has as expected no impact on the combination of n-best lists of the same size (see Table 2), however it shows an improvement of BLEU +0.5 for the combination with the 1st-best from system B. The smoothing seems to have no significant impact for this dataset, but different smoothing factors will be investigated in the future.

4 Workshop Results

To train our language models and word lexica we only used provided data. Therefore we excluded systems from the combination, which were to our knowledge using unrestricted training data (google). We did not include any contrastive systems.

We trained the statistical word lexica on the parallel data provided for each language pair[1]. For each combination we used two language models, a 1.2 giga-word 3-gram language model, trained on the provided monolingual English data and a 4-gram language model trained on the English part of the parallel training data of the respective languages. We used the SRILM toolkit (Stolcke, 2002) for training.

For each of the three language pairs we submitted a combination that used the plain version of the n-gram agreement feature as well as one using the normalized smoothed version.

The provided system combination development set, which we used for tuning our feature weights, was the same for all language pairs, 502 sentences with only one reference.

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 results for TER and METEOR are not very meaningful here and only reported for completeness.

4.1 French-English

14 systems were submitted to the restricted data track for the French-English translation task. The scores on the combination development set range from BLEU 27.56 to 15.09 (case insensitive evaluation).

We received n-best lists from five systems, a 300-best, a 200-best two 100-best and one 10-best list. We included up to 100 hypotheses per system in our joint n-best list.

For our workshop submission we combined the top nine systems with the last system scoring 24.23 as well as all 14 systems. Comparing the results for the two combinations of all 14 systems (see Table 3), the one with the sub-list normalization for the n-gram agreement feature gains +0.8 BLEU on unseen data compared to the one without normalization.

system	dev	test	TER	Meteor
best single	27.56	26.88	56.32	52.68
top 9 a_{smooth}	29.85	28.07	55.23	53.90
all 14 a_{smooth}	30.39	28.46	55.12	54.35
all 14	29.49	27.65	55.41	53.74

Table 3: French-English Results: BLEU

Our system combination via hypothesis selection could improve the translation quality by +1.6 BLEU on the unseen test set compared to the best single system.

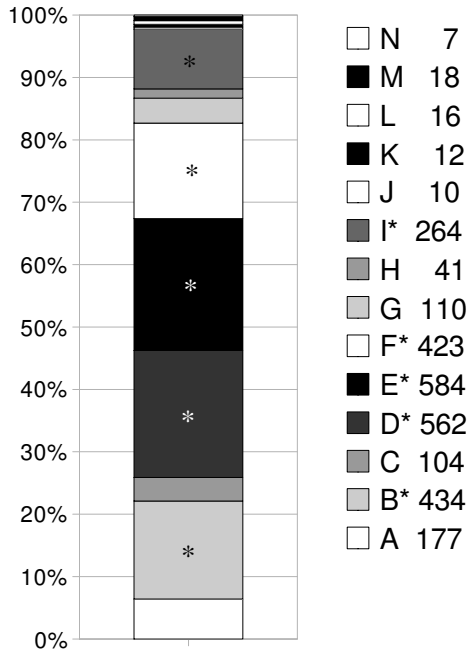

Figure 1: Contributions of the individual systems to the final translation.

Figure 1 shows, how many hypotheses were contributed by the individual systems to the final translation (unseen data). The systems A to N are ordered by their BLEU score on the development set. The systems which provided n-best lists, marked with a star in the diagram, clearly dominate the selection. The low scoring systems contribute very little as expected.

4.2 German-English

14 systems were submitted to the restricted data track for the German-English translation task. The scores on the combination development set range

[1] http://www.statmt.org/wmt09/translation-task.html#training

from BLEU 27.56 to 7 (case insensitive evaluation). The two lowest scoring systems at BLEU 11 and 7 were so far from the rest of the systems that we decided to exclude them, assuming an error had occurred.

Within the remaining 12 submissions were four n-best lists, three 100-best and one 10-best.

For our submissions we combined the top seven systems between BLEU 22.91 and 20.24 as well as the top 12 systems where the last one of those was scoring BLEU 16.00 on the development set. For this language pair the combination with the normalized n-gram agreement also outperforms the one without by +0.8 BLEU (see Table 4).

system	dev	test	TER	Meteor
best single	22.91	21.03	61.87	47.96
top 7 a_{smooth}	25.13	22.86	60.73	49.71
top 12 a_{smooth}	25.32	22.98	60.72	50.01
top 12	25.12	22.20	60.95	49.33

Table 4: German-English Results: BLEU

Our system combination via hypothesis selection could improve translation quality by +1.95 BLEU on the unseen test set over the best single system.

4.3 Hungarian-English

Only three systems were submitted for the Hungarian-English translation task. Scores on the combination development set ranged from BLEU 13.63 to 10.04 (case insensitive evaluation). Only the top system provided an n-best list. We used 100-best hypotheses.

system	dev	test	TER	Meteor
best single	13.63	12.73	68.75	36.76
3 sys a_{smooth}	14.98	13.74	72.34	38.20
3 sys	14.14	13.18	74.29	37.52

Table 5: Hungarian-English Results: BLEU

We submitted combinations of the three systems by using the modified smoothed n-gram agreement feature and the plain version of the n-gram agreement feature. Here also the normalized version of the feature gives an improvement of +0.56 BLEU with an overall improvement of +1.0 BLEU over the best single system (see Table 5).

5 Summary

It is beneficial to include more systems, even if they are more than 7 points BLEU behind the best system, as the comparison to the combinations with fewer systems shows.

In the mixed size data situation of the workshop the modified feature shows a clear improvement for all three language pairs. Different smoothing factors should be investigated for these data sets in the future.

Acknowledgments

We would like to thank the participants in the WMT'09 workshop shared translation task for providing their data, especially n-best lists.

References

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.

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.

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.