Incremental Hypothesis Alignment for Building Confusion Networks with Application to Machine Translation System Combination

Antti-Veikko I. Rosti and Bing Zhang and Spyros Matsoukas and Richard Schwartz
BBN Technologies, 10 Moulton Street, Cambridge, MA 02138
{arosti,bzhang,smatsouk,schwartz}@bbn.com

Abstract

Confusion network decoding has been the most successful approach in combining outputs from multiple machine translation (MT) systems in the recent DARPA GALE and NIST Open MT evaluations. Due to the varying word order between outputs from different MT systems, the hypothesis alignment presents the biggest challenge in confusion network decoding. This paper describes an incremental alignment method to build confusion networks based on the translation edit rate (TER) algorithm. This new algorithm yields significant BLEU score improvements over other recent alignment methods on the GALE test sets and was used in BBN’s submission to the WMT08 shared translation task.

1 Introduction

Confusion network decoding has been applied in combining outputs from multiple machine translation systems. The earliest approach in (Bangalore et al., 2001) used edit distance based multiple string alignment (MSA) (Durbin et al., 1988) to build the confusion networks. The recent approaches used pair-wise alignment algorithms based on symmetric alignments from a HMM alignment model (Matusov et al., 2006) or edit distance alignments allowing shifts (Rosti et al., 2007). The alignment method described in this paper extends the latter by incrementally aligning the hypotheses as in MSA but also allowing shifts as in the TER alignment.

The confusion networks are built around a “skeleton” hypothesis. The skeleton hypothesis defines the word order of the decoding output. Usually, the 1-best hypotheses from each system are considered as possible skeletons. Using the pair-wise hypothesis alignment, the confusion networks are built in two steps. First, all hypotheses are aligned against the skeleton independently. Second, the confusion networks are created from the union of these alignments. The incremental hypothesis alignment algorithm combines these two steps. All words from the previously aligned hypotheses are available, even if not present in the skeleton hypothesis, when aligning the following hypotheses. As in (Rosti et al., 2007), confusion networks built around all skeletons are joined into a lattice which is expanded and re-scored with language models. System weights and language model weights are tuned to optimize the quality of the decoding output on a development set.

This paper is organized as follows. The incremental TER alignment algorithm is described in Section 2. Experimental evaluation comparing the incremental and pair-wise alignment methods are presented in Section 3 along with results on the WMT08 Europarl test sets. Conclusions and future work are presented in Section 4.

2 Incremental TER Alignment

The incremental hypothesis alignment is based on an extension of the TER algorithm (Snover et al., 2006). The extension allows using a confusion network as the reference. First, the algorithm finds the minimum edit distance between the hypothesis and the reference network by considering all word arcs between two consecutive nodes in the reference network as possible matches for a hypothesis word at
that position. Second, shifts of blocks of words that
have an exact match somewhere else in the network
are tried in order to find a new hypothesis word or-
der with a lower TER. Each shifted block is con-
sidered a single edit. These two steps are executed
iteratively as a greedy search. The final alignment
between the re-ordered hypothesis and the reference
network may include matches, substitutions, dele-
tions, and insertions.

The confusion networks are built by creating a
simple confusion network from the skeleton hypoth-
esis. If the skeleton hypothesis has \( N \) words, the
initial network has \( N \) arcs and \( N + 1 \) nodes. Each
arc has a set of system specific confidence scores.
The score for the skeleton system is set to \( 1/2 \) and
the confidences for other systems are set to zeros.
For each non-skeleton hypothesis, a TER alignment
against the current network is executed as described
above. Each match found will increase the system
specific word arc confidence by \( 1/(1 + k) \) where \( k \)
is the rank of the hypothesis in that system’s \( N \)-best
list. Each substitution will generate a new word arc
at the corresponding position in the network. The
word arc confidence for the system is set to \( 1/(1 + k) \)
and the confidences for other systems are set to ze-
ros. Each deletion will generate a new NULL word
arc unless one exists at the corresponding position
in the network. The NULL word arc confidences are
adjusted as in the case of a match or a substitution
depending on whether the NULL word arc exists or
not. Finally, each insertion will generate a new node
and two word arcs at the corresponding position in
the network. The first word arc will have the in-
serted word with the confidence set as in the case
of a substitution and the second word arc will have
a NULL word with confidences set by assuming all
previously aligned hypotheses and the skeleton gen-
erated the NULL word arc.

After all hypotheses have been added into the con-
fusion network, the system specific word arc confi-
dences are scaled to sum to one over all arcs between
each set of two consecutive nodes. Other scores for
the word arc are set as in (Rosti et al., 2007).

2.1 Benefits over Pair-Wise TER Alignment

The incremental hypothesis alignment guarantees
that insertions between a hypothesis and the cur-
cent confusion network are always considered when
aligning the following hypotheses. This is not the
case in any pair-wise hypothesis alignment algo-
rum. During the pair-wise hypothesis alignment,
an identical word in two hypotheses may be aligned
as an insertion or a substitution in a different posi-
tion with respect to the skeleton. This will result in
undesirable repetition and lower confidence for that
word in the final confusion network. Also, multiple
insertions are not handled implicitly.

For example, three hypotheses “I like balloons”,
“I like big blue balloons”, and “I like blue kites”
might be aligned by the pair-wise alignment, assum-
ing the first as the skeleton, as follows:

\[
\begin{array}{c}
\text{I like } \text{NULL } \text{balloons } \text{NULL} \\
\text{I like } \text{big blue } \text{balloons } \text{NULL} \\
\end{array}
\]

which results in the confusion network shown in
Figure 1. The number of hypotheses proposing each
word is shown in parentheses. The alignment be-
 tween the skeleton and the second hypothesis has
two consecutive insertions “big blue” which are not
available for matching when the third hypothesis is
aligned against the skeleton. Therefore, the word
“blue” appears twice in the confusion network. If
many hypotheses have multiple insertions at the
same location with respect to the skeleton, they have
to be treated as phrases or a secondary alignment
process has to be applied.

Assuming the same hypotheses as above, the in-
cremental hypothesis alignment may yield the fol-
lowing alignment:
| System          | TER  | BLEU | MTR  |
|-----------------|------|------|------|
| worst           | 53.26| 33.00| 63.15|
| best            | 42.30| 48.52| 67.71|
| syscomb pw      | 39.85| 52.00| 68.73|
| syscomb giza    | 40.01| 52.24| 68.68|
| syscomb inc     | 39.25| 52.73| 68.97|
| oracle          | 21.68| 64.14| 78.18|

Table 1: Results on the Arabic GALE Phase 2 system combination tuning set with four reference translations.

| System          | TER  | BLEU | MTR  |
|-----------------|------|------|------|
| worst           | 59.09| 20.74| 57.24|
| best            | 48.18| 31.46| 62.61|
| syscomb pw      | 46.31| 33.02| 63.18|
| syscomb giza    | 46.03| 33.39| 63.21|
| syscomb inc     | 45.45| 33.90| 63.45|
| oracle          | 27.53| 49.10| 71.81|

Table 2: Results on the Arabic GALE Phase 2 evaluation set with one reference translation.

I like NULL NULL balloons
I like big blue balloons
I like NULL blue kites

which results in the confusion network shown in Figure 2. In this case the word “blue” is available for matching when the third hypothesis is aligned. It should be noted that the final confusion network depends on the order in which the hypotheses are added. The experiments so far have indicated that different alignment order does not have a significant influence on the final combination results as measured by the automatic evaluation metrics. Usually, aligning the system outputs in the decreasing order of their TER scores on the development set yields the best scores.

### 2.2 Confusion Network Oracle

The extended TER algorithm can also be used to estimate an oracle TER in a confusion network by aligning the reference translations against the confusion network. The oracle hypotheses can be extracted by finding a path with the maximum number of matches. These hypotheses give a lower bound on the TER score for the hypotheses which can be generated from the confusion networks.

### 3 Experimental Evaluation

The quality of the final combination output depends on many factors. Combining very similar outputs does not yield as good gains as combining outputs from diverse systems. It is also important that the development set used to tune the combination weights is as similar to the evaluation set as possible. This development set should be different from the one used to tune the individual systems to avoid bias toward any system that may be over-tuned. Due to the tight schedule for the WMT08, there was no time to experiment with many configurations. As more extensive experiments have been conducted in the context of the DARPA GALE program, results on the Arabic GALE Phase 2 evaluation setup are first presented. The translation quality is measured by three MT evaluation metrics: TER (Snover et al., 2006), BLEU (Papineni et al., 2002), and METEOR (Lavie and Agarwal, 2007).

### 3.1 Results on Arabic GALE Outputs

For the Arabic GALE Phase 2 evaluation, nine systems were combined. Five systems were phrase-based, two hierarchical, one syntax-based, and one rule-based. All statistical systems were trained on common parallel data, tuned on a common genre specific development set, and a common English tokenization was used. The English bi-gram and 5-gram language models used in the system combination were trained on about 7 billion words of English text. Three iterations of bi-gram decoding weight tuning were performed followed by one iteration of 5-gram re-scoring weight tuning. All weights were tuned to minimize the sum of TER and 1-BLEU. The final 1-best outputs were true-cased and detokenized before scoring.

The results on the newswire system combination development set and the GALE Phase 2 evaluation set are shown in Tables 1 and 2. The first two rows show the worst and best scores from the individual systems. The scores may be from different systems as the best performing system in terms of TER was not necessarily the best performing system in terms of the other metrics. The following three rows show the scores of three combination outputs where the only difference was the hypothesis alignment method. The first, syscomb pw, corresponds...
| BLEU          | System de-en fr-en |
|--------------|--------------------|
| worst        | 11.84 16.31        |
| best         | 28.30 33.13        |
| syscomb      | 29.05 33.63        |

Table 3: NIST BLEU scores on the German-English (de-en) and French-English (fr-en) Europarl test2008 set.

February 1, 2008 (Updated on 2 March 2008)

the pair-wise TER alignment described in (Rosti et al., 2007). The second, syscomb giza, corresponds to the pair-wise symmetric HMM alignments from GIZA++ described in (Matusov et al., 2006). The third, syscomb inc, corresponds to the incremental TER alignment presented in this paper. Finally, oracle corresponds to an estimate of the lower bound on the translation quality obtained by extracting the TER oracle output from the confusion networks generated by the incremental TER alignment. It is unlikely that there exists a set of weights that would yield the oracle output after decoding, though. The incremental TER alignment yields significant improvements over all individual systems and the combination outputs using the pair-wise alignment methods.

3.2 Results on WMT08 Europarl Outputs

On the WMT08 shared translation task, translations for two language pairs and two tasks were provided for the system combination experiments. Twelve systems participated in the German-English and fourteen in the French-English translation tasks. The translations of the Europarl test (test2008) were provided as the development set outputs and the translations of the News test (newstest2008) were provided as the evaluation set outputs. An English bi-gram, 4-gram, and true-caser language models were trained by using all English text available for the WMT08 shared task, including Europarl monolingual and news commentary parallel training sets. The outputs were tokenized and lower-cased before combination, and the final combination output was true-cased and detokenized.

The results on the Europarl test set for both language pairs are shown in table 3. The first two rows have the NIST BLEU scores of the worst and the best individual systems. The last row, syscomb, corresponds to the system combination using the incremental TER alignment. The improvements in the NIST BLEU scores are fairly modest which is probably due to low diversity of the system outputs. It is also unlikely that these weights are optimal for the out-of-domain News test set outputs.

4 Conclusions

This paper describes a novel hypothesis alignment algorithm for building confusion networks from multiple machine translation system outputs. The algorithm yields significant improvements on the Arabic GALE evaluation set outputs and was used in BBN’s submission to the WMT08 shared translation task. The hypothesis alignment may benefit from using stemming and synonymy in matching words. Also, special handling of punctuation may improve the alignment further. The future work will investigate the influence of better alignment to the final combination outputs.

Acknowledgments

This work was supported by DARPA/IPTO Contract No. HR0011-06-C-0022 under the GALE program.

References

S. Bangalore, G. Bordel, and G. Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In Proc. ASRU, pages 351–354.
R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. 1988. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge Univ. Press.
A. Lavie and A. Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proc. ACL/WMT, pages 228–231.
E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In Proc. EACL, pages 33–40.
K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. ACL, pages 311–318.
A.-V.I. Rosti, S. Matsoukas, and R. Schwartz. 2007. Improved word-level system combination for machine translation. In Proc. ACL 2007, pages 312–319.
M. Snoover, B. Dorr, R. Schwartz, L. Micciula, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In Proc. AMIA, pages 223–231.