# Incremental Hypothesis Alignment with Flexible Matching for Building Confusion Networks: BBN System Description for WMT09 System Combination Task

**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

This paper describes the incremental hypothesis alignment algorithm used in the BBN submissions to the WMT09 system combination task. The alignment algorithm used a sentence specific alignment order, flexible matching, and new shift heuristics. These refinements yield more compact confusion networks compared to using the pair-wise or incremental TER alignment algorithms. This should reduce the number of spurious insertions in the system combination output and the system combination weight tuning converges faster. System combination experiments on the WMT09 test sets from five source languages to English are presented. The best BLEU scores were achieved by combing the English outputs of three systems from all five source languages.

## 1 Introduction

Machine translation (MT) systems have different strengths and weaknesses which can be exploited by system combination methods resulting in an output with a better performance than any individual MT system output as measured by automatic evaluation metrics. Confusion network decoding has become the most popular approach to MT system combination. The first confusion network decoding method (Bangalore et al., 2001) was based on multiple string alignment (MSA) (Durbin et al., 1988) borrowed from biological sequence analysis. However, MSA does not allow re-ordering. The translation edit rate (TER) (Snover et al., 2006) produces an alignment between two strings and allows shifts of blocks of words. The availability of the TER software has made it easy to build a high performance system combination baseline (Rosti et al., 2007).

The pair-wise TER alignment originally described by Sim et al. (2007) has various limitations. First, the hypotheses are aligned independently against the skeleton which determines the word order of the output. The same word from two different hypotheses may be inserted in different positions w.r.t. the skeleton and multiple insertions require special handling. Rosti et al. (2008) described an incremental TER alignment to mitigate these problems. The incremental TER alignment used a global order in which the hypotheses were aligned. Second, the TER software matches words with identical surface strings. The pair-wise alignment methods proposed by Ayan et al. (2008), He et al. (2008), and Matusov et al. (2006) are able to match also synonyms and words with identical stems. Third, the TER software uses a set of heuristics which is not always optimal in determining the block shifts. Karakos et al. (2008) proposed using inversion transduction grammars to produce different pair-wise alignments.

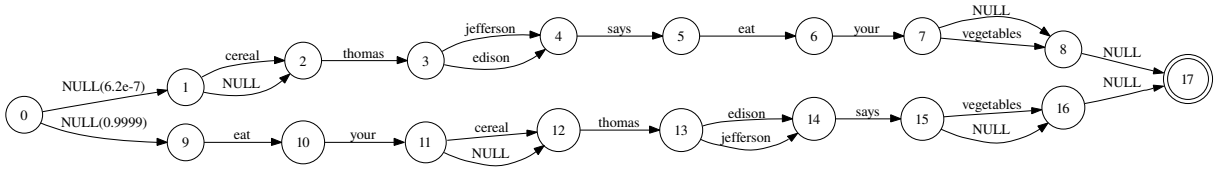
This paper is organized as follows. A refined incremental alignment algorithm is described in Section 2. Experimental evaluation comparing the pair-wise and incremental TER alignment algorithms with the refined alignment algorithm on WMT09 system combination task is presented in Section 3. Conclusions and future work are presented in Section 4.

## 2 Incremental Hypothesis Alignment with Flexible Matching

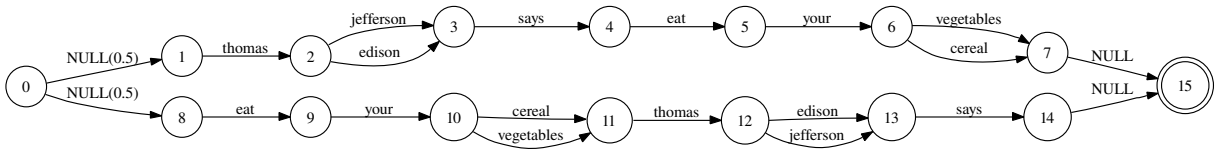
### 2.1 Sentence Specific Alignment Order

Rosti et al. (2008) proposed incremental hypothesis alignment using a system specific order. This is not likely to be optimal since one MT system may have better output on one sentence and worse on another. More principled approach is similar to MSA where the order is determined by the edit distance of the hypothesis from the network for

*Proceedings of the Fourth Workshop on Statistical Machine Translation* , pages 61–65,
Athens, Greece, 30 March – 31 March 2009. 




(a) Alignment using the standard TER shift heuristics.




(b) Alignment using the modified shift heuristics.

Figure 1: Combined confusion networks using different shift heuristics. The initial NULL arcs include the prior probability estimates in parentheses.

each sentence. The TER scores of the remaining unaligned hypotheses using the current network as the reference are computed. The hypothesis with the lowest edit cost w.r.t. the network is aligned. Given $N$ systems, this increases the number of alignments performed from $N$ to $0.5(N^2 - N)$.

### 2.2 Flexible Matching

The TER software assigns a zero cost for matching tokens and a cost of one for all errors including insertions, deletions, substitutions, and block shifts. Ayan et al. (2008) modified the TER software to consider substitutions of synonyms with a reduced cost. Recently, Snover et al. (2009) extended the TER algorithm in a similar fashion to produce a new evaluation metric, TER plus (TERp), which allows tuning of the edit costs in order to maximize correlation with human judgment. The incremental alignment with flexible matching uses WordNet (Fellbaum, 1998) to find all possible synonyms and words with identical stems in a set of hypotheses. Substitutions involving synonyms and words with identical stems are considered with a reduced cost of 0.2.

### 2.3 Modified Shift Heuristics

The TER is computed by trying shifts of blocks of words that have an exact match somewhere else in the reference in order to find a re-ordering of the hypothesis with a lower edit distance to the reference. Karakos et al. (2008) showed that the shift heuristics in TER do not always yield an optimal alignment. Their example used the following two hypotheses:

1. thomas jefferson says eat your vegetables

2. eat your cereal thomas edison says

A system combination lattice using TER alignment is shown in Figure 1(a). The blocks "eat your" are shifted when building both confusion networks. Using the second hypothesis as the skeleton seems to give a better alignment. The lower number of edits also results in a higher skeleton prior shown between nodes 0 and 9. There are obviously some undesirable paths through the lattice but it is likely that a language model will give a higher score to the reasonable hypotheses.

Since the flexible matching allows substitutions with a reduced cost, the standard TER shift heuristics have to be modified. A block of words may have some words with identical matches and other words with synonym matches. In TERp, synonym and stem matches are considered as exact matches for the block shifts, otherwise the TER shift constraints are used. In the flexible matching, the shift heuristics were modified to allow any block shifts

that do not increase the edit cost. A system combination lattice using the modified shift heuristics is shown in Figure 1(b). The optimal shifts of blocks "eat your cereal" and "eat your vegetables" were found and both networks received equal skeleton priors. TERp would yield this alignment only if these blocks appear in the paraphrase table or if "cereal" and "vegetables" are considered synonyms. This example is artificial and does not guarantee that optimal shifts are always found.

## 3 Experimental Evaluation

System combination experiments combining the English WMT09 translation task outputs were performed. A total of 96 English outputs were provided including primary, contrastive, and $N$-best outputs. Only the primary 1-best outputs were combined due to time constraints. The numbers of primary systems per source language were: 3 for Czech, 15 for German, 9 for Spanish, 15 for French, and 3 for Hungarian. The English bigram and 5-gram language models were interpolated from four LM components trained on the English monolingual Europarl (45M tokens) and News (510M tokens) corpora, and the English sides of the News Commentary (2M tokens) and GigaFrEn (683M tokens) parallel corpora. The interpolation weights were tuned to minimize perplexity on `news-dev2009` set. The system combination weights – one for each system, LM weight, and word and NULL insertion penalties – were tuned to maximize the BLEU (Papineni et al., 2002) score on the tuning set (`newssyscomb2009`). Since the system combination was performed on tokenized and lower cased outputs, a trigram-based true caser was trained on all News training data. The tuning may be summarized as follows:

1. Tokenize and lower case the outputs;
2. Align hypotheses incrementally using each output as a skeleton;
3. Join the confusion networks into a lattice with skeleton specific prior estimates;
4. Extract a 300-best list from the lattice given the current weights;
5. Merge the 300-best list with the hypotheses from the previous iteration;
6. Tune new weights given the current merged $N$-best list;
7. Iterate 4-6 three times;
8. Extract a 300-best list from the lattice given the best decoding weights and re-score hypotheses with a 5-gram;
9. Tune re-scoring weights given the final 300-best list;
10. Extract 1-best hypotheses from the 300-best list given the best re-scoring weights, re-case, and detokenize.

After tuning the system combination weights, the outputs on a test set may be combined using the same steps excluding 4-7 and 9. The hypothesis scores and tuning are identical to the setup used in (Rosti et al., 2007).

Case insensitive TER and BLEU scores for the combination outputs using the pair-wise and incremental TER alignment as well as the flexible alignment on the tuning (dev) and test sets are shown in Table 1. Only case insensitive scores are reported since the re-casers used by different systems are very different and some are trained using larger resources than provided for WMT09. The scores of the worst and best individual system outputs are also shown. The best and worst TER and BLEU scores are not necessarily from the same system output. Both `incremental` and `flexible` alignments used sentence specific alignment order. Combinations using the incremental and flexible hypothesis alignment algorithms consistently outperform the ones using the pair-wise TER alignment. The flexible alignment is slightly better than the incremental alignment on Czech, Spanish, and Hungarian, and significantly better on French to English test set scores.

Since the test sets for each language pair consist of translations of the same documents, it is possible to combine outputs from many source languages to English. There were a total of 46 English primary 1-best system outputs. Using all 46 outputs would have required too much memory in tuning, so a subset of 11 outputs was chosen. The 11 outputs consist of `google`, `uedin`, and `uka` outputs on all languages. Case insensitive TER and BLEU scores for the `xx-en` combination are shown in Table 2. In addition to `incremental` and `flexible` alignment methods which used sentence specific alignment order, scores for incremental TER alignment with a fixed alignment order used in the BBN submissions to WMT08

| **dev** | cz-en | | de-en | | es-en | | fr-en | | hu-en | |
|---|---|---|---|---|---|---|---|---|---|---|
| System | TER | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER | BLEU |
| worst | 67.30 | 17.63 | 82.01 | 6.83 | 65.64 | 19.74 | 69.19 | 15.21 | 78.70 | 10.33 |
| best | 58.16 | 23.12 | 57.24 | 23.20 | 53.02 | 29.48 | 49.78 | 32.27 | 66.77 | 13.59 |
| pairwise | 59.60 | 24.01 | 56.35 | 26.04 | 53.11 | 29.49 | 51.03 | 31.65 | 69.58 | 14.60 |
| incremental | 59.22 | 24.31 | 55.73 | 26.73 | 53.05 | 29.72 | 50.72 | 32.09 | 70.15 | 14.85 |
| flexible | 59.38 | 24.18 | 55.51 | 26.71 | 52.62 | 30.24 | 50.22 | 32.58 | 69.83 | 14.88 |

| **test** | cz-en | | de-en | | es-en | | fr-en | | hu-en | |
|---|---|---|---|---|---|---|---|---|---|---|
| System | TER | BLEU | TER | BLEU | TER | BLEU | TER | BLEU | TER | BLEU |
| worst | 67.74 | 16.37 | 82.39 | 6.81 | 65.44 | 19.04 | 71.44 | 14.49 | 81.21 | 9.90 |
| best | 59.53 | 21.18 | 59.41 | 21.30 | 53.34 | 28.69 | 51.33 | 31.14 | 68.32 | 12.75 |
| pairwise | 61.02 | 21.25 | 58.75 | 23.41 | 53.65 | 28.15 | 53.17 | 29.83 | 71.50 | 13.39 |
| incremental | 60.63 | 21.67 | 58.13 | 23.96 | 53.47 | 28.38 | 52.51 | 30.45 | 71.69 | 13.60 |
| flexible | 60.34 | 21.87 | 58.05 | 23.86 | 53.13 | 28.57 | 51.98 | 31.30 | 71.17 | 13.84 |

Table 1: Case insensitive TER and BLEU scores on newssyscomb2009 (dev) and newstest2009 (test) for five source languages.

(Rosti et al., 2008) are marked as incr-wmt08. The sentence specific alignment order yields about a half BLEU point gain on the tuning set and a one BLEU point gain on the test set. All system combination experiments yield very good BLEU gains on both sets. The scores are also significantly higher than any combination from a single source language. This shows that the outputs from different source languages are likely to be more diverse than outputs from different MT systems on a single language pair. The combination is not guaranteed to be the best possible as the set of outputs was chosen arbitrarily.

The compactness of the confusion networks may be measured by the average number of nodes and arcs per segment. All xx-en confusion networks for newssyscomb2009 and newstest2009 after the incremental TER alignment had on average 44.5 nodes and 112.7 arcs per segment. After the flexible hypothesis alignment, there were on average 41.1 nodes and 104.6 arcs per segment. The number of NULL word arcs may also be indicative of the alignment quality. The flexible hypothesis alignment reduced the average number of NULL word arcs from 29.0 to 24.8 per segment. The rate of convergence in the $N$-best list based iterative tuning may be monitored by the number of new hypotheses in the merged $N$-best lists from iteration to iteration. By the third tuning iteration, there were 10% fewer new hypotheses in the merged $N$-best list when using the flexible hypothesis alignment.

| xx-en | **dev** | | **test** | |
|---|---|---|---|---|
| System | TER | BLEU | TER | BLEU |
| worst | 74.21 | 12.80 | 75.84 | 12.05 |
| best | 49.78 | 32.27 | 51.33 | 31.14 |
| pairwise | 46.10 | 35.95 | 47.77 | 33.53 |
| incr-wmt08 | 44.58 | 36.84 | 46.60 | 33.61 |
| incremental | 44.59 | 37.30 | 46.42 | 34.61 |
| flexible | 44.54 | 37.38 | 45.82 | 34.48 |

Table 2: Case insensitive TER and BLEU scores on newssyscomb2009 (dev) and newstest2009 (test) for xx-en combination.

## 4 Conclusions

This paper described a refined incremental hypothesis alignment algorithm used in the BBN submissions to the WMT09 system combination task. The new features included sentence specific alignment order, flexible matching, and modified shift heuristics. The refinements yield more compact confusion networks which should allow fewer spurious insertions in the output and faster convergence in tuning. The future work will investigate tunable edit costs and methods to choose an optimal subset of outputs for combination.

## Acknowledgments

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

## References


Necip Fazil Ayan, Jing Zheng, and Wen Wang. 2008. Improving alignments for better confusion networks for combining machine translation systems. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 33–40.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 351–354.

Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1988. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-HMM-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 98–107.

Damianos Karakos, Jason Eisner, Sanjeev Khundanpur, and Markus Dreyer. 2008. Machine translation system combination using ITG-based alignments. In *Proceedings of ACL-08: HLT*, pages 81–84.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Antti-Veikko I. Rosti, Spyros Matsoukas, and Richard Schwartz. 2007. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 312–319.

Antti-Veikko I. 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.

Khe Chai Sim, William J. Byrne, Mark J.F. Gales, Hichem Sahbi, and Phil C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 105–108.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciula, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*.