Review of Hypothesis Alignment Algorithms for MT System Combination via Confusion Network Decoding

Antti-Veikko I. Rosti[a]*, **Xiaodong He**[b], **Damianos Karakos**[c], **Gregor Leusch**[d]†, **Yuan Cao**[c], **Markus Freitag**[e], **Spyros Matsoukas**[f], **Hermann Ney**[e], **Jason R. Smith**[c] and **Bing Zhang**[f]

[a]Apple Inc., Cupertino, CA 95014
arosti@apple.com

[b]Microsoft Research, Redmond, WA 98052
xiaohe@microsoft.com

[c]Johns Hopkins University, Baltimore, MD 21218
{damianos,yuan.cao,jrsmith}@jhu.edu

[d]SAIC, Monheimsallee 22, D-52062 Aachen, Germany
gregor.leusch@saic.com

[e]RWTH Aachen University, D-52056 Aachen, Germany
{freitag,ney}@cs.rwth-aachen.de

[f]Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138
{smatsouk,bzhang}@bbn.com

Abstract

Confusion network decoding has proven to be one of the most successful approaches to machine translation system combination. The hypothesis alignment algorithm is a crucial part of building the confusion networks and many alternatives have been proposed in the literature. This paper describes a systematic comparison of five well known hypothesis alignment algorithms for MT system combination via confusion network decoding. Controlled experiments using identical pre-processing, decoding, and weight tuning methods on standard system combination evaluation sets are presented. Translation quality is assessed using case insensitive BLEU scores and bootstrapping is used to establish statistical significance of the score differences. All aligners yield significant BLEU score gains over the best individual system included in the combination. Incremental indirect hidden Markov model and a novel incremental inversion transduction grammar with flexible matching consistently yield the best translation quality, though keeping all things equal, the differences between aligners are relatively small.

1 Introduction

Current machine translation (MT) systems are based on different paradigms, such as rule-based, phrase-based, hierarchical, and syntax-based. Due to the complexity of the problem, systems make various assumptions at different levels of processing and modeling. Many of these assumptions may be suboptimal and complementary. The complementary information in the outputs from multiple MT systems may be exploited by system combination. Availability of multiple system outputs within the DARPA GALE program as well as NIST Open MT and Workshop on Statistical Machine Translation evaluations has led to extensive research in combining the strengths of diverse MT systems, resulting in significant gains in translation quality.

System combination methods proposed in the literature can be roughly divided into three categories: (i) hypothesis selection (Rosti et al., 2007b; Hildebrand and Vogel, 2008), (ii) re-decoding (Frederking and Nirenburg, 1994; Jayaraman and Lavie, 2005; Rosti et al., 2007b; He and Toutanova, 2009; Devlin et al., 2011), and (iii) confusion network decoding. Confusion network decoding has proven to be the most popular as it does not require deep N-best lists[1] and operates on the surface strings. It has

*The work reported in this paper was carried out while the authors were at Raytheon BBN Technologies and †RWTH Aachen University.

[1] N-best lists of around $N = 10$ have been used in confusion network decoding yielding small gains over using 1-best

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 191–199,
Montréal, Canada, June 7-8, 2012.

also been shown to be very successful in combining speech recognition outputs (Fiscus, 1997; Mangu et al., 2000). The first application of confusion network decoding in MT system combination appeared in (Bangalore et al., 2001) where a multiple string alignment (MSA), made popular in biological sequence analysis, was applied to the MT system outputs. Matusov et al. (2006) proposed an alignment based on GIZA++ Toolkit which introduced word reordering not present in MSA, and Sim et al. (2007) used the alignments produced by the translation edit rate (TER) (Snover et al., 2006) scoring. Extensions of the last two are included in this study together with alignments based on hidden Markov model (HMM) (Vogel et al., 1996) and inversion transduction grammars (ITG) (Wu, 1997).

System combinations produced via confusion network decoding using different hypothesis alignment algorithms have been entered into open evaluations, most recently in 2011 Workshop on Statistical Machine Translation (Callison-Burch et al., 2011). However, there has not been a comparison of the most popular hypothesis alignment algorithms using the same sets of MT system outputs and otherwise identical combination pipelines. This paper attempts to systematically compare the quality of five hypothesis alignment algorithms. Alignments were produced for the same system outputs from three common test sets used in the 2009 NIST Open MT Evaluation and the 2011 Workshop on Statistical Machine Translation. Identical pre-processing, decoding, and weight tuning algorithms were used to quantitatively evaluate the alignment quality. Case insensitive BLEU score (Papineni et al., 2002) was used as the translation quality metric.

2 Confusion Network Decoding

A confusion network is a linear graph where all paths visit all nodes. Two consecutive nodes may be connected by one or more arcs. Given the arcs represent words in hypotheses, multiple arcs connecting two consecutive nodes can be viewed as alternative words in that position of a set of hypotheses encoded by the network. A special NULL token represents a skipped word and will not appear in the system combination output. For example, three hypotheses "twelve big cars", "twelve cars", and "dozen cars" may be aligned as follows:

twelve	big	blue	cars
twelve	NULL	NULL	cars
dozen	NULL	blue	cars

This alignment may be represented compactly as the confusion network in Figure 1 which encodes a total of eight unique hypotheses.

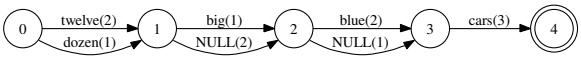

Figure 1: Confusion network from three strings "twelve big blue cars", "twelve cars", and "dozen blue cars" using the first as the skeleton. The numbers in parentheses represent counts of words aligned to the corresponding arc.

Building confusion networks from multiple machine translation system outputs has two main problems. First, one output has to be chosen as the skeleton hypothesis which defines the final word order of the system combination output. Second, MT system outputs may have very different word orders which complicates the alignment process. For skeleton selection, Sim et al. (2007) proposed choosing the output closest to all other hypotheses when using each as the reference string in TER. Alternatively, Matusov et al. (2006) proposed leaving the decision to decoding time by connecting networks built using each output as a skeleton into a large lattice. The subnetworks in the latter approach may be weighted by prior probabilities estimated from the alignment statistics (Rosti et al., 2007a). Since different alignment algorithm produce different statistics and the gain from the weights is relatively small (Rosti et al., 2011), weights for the subnetworks were not used in this work. The hypothesis alignment algorithms used in this work are briefly described in the following section.

The confusion networks in this work were represented in a text lattice format shown in Figure 2. Each line corresponds to an arc, where J is the arc index, S is the start node index, E is the end node index, SC is the score vector, and W is the word label. The score vector has as many elements as there are input systems. The elements correspond to each system and indicate whether a word from a particular

outputs (Rosti et al., 2011).

```
J=0 S=0 E=1 SC=(1,1,0) W=twelve
J=1 S=0 E=1 SC=(0,0,1) W=dozen
J=2 S=1 E=2 SC=(1,0,0) W=big
J=3 S=1 E=2 SC=(0,1,1) W=NULL
J=4 S=2 E=3 SC=(1,0,1) W=blue
J=5 S=2 E=3 SC=(0,1,0) W=NULL
J=6 S=3 E=4 SC=(1,1,1) W=cars
```

Figure 2: A lattice in text format representing the confusion network in Figure 1. J is the arc index, S and E are the start and end node indexes, SC is a vector of arc scores, and W is the word label.

system was aligned to a given link[2]. These may be viewed as system specific word confidences, which are binary when aligning 1-best system outputs. If no word from a hypothesis is aligned to a given link, a NULL word token is generated provided one does not already exist, and the corresponding element in the NULL word token is set to one. The system specific word scores are kept separate in order to exploit system weights in decoding. Given system weights w_n, which sum to one, and system specific word scores s_{nj} for each arc j (the SC elements), the weighted word scores are defined as:

$$s_j = \sum_{n=1}^{N_s} w_n s_{nj} \quad (1)$$

where N_s is the number of input systems. The hypothesis score is defined as the sum of the log-word-scores along the path, which is linearly interpolated with a logarithm of the language model (LM) score and a non-NULL word count:

$$S(E|F) = \sum_{j \in \mathcal{J}(E)} \log s_j + \gamma S_{LM}(E) + \delta N_w(E) \quad (2)$$

where $\mathcal{J}(E)$ is the sequence of arcs generating the hypothesis E for the source sentence F, $S_{LM}(E)$ is the LM score, and $N_w(E)$ is the number of non-NULL words. The set of weights $\theta = \{w_1, \ldots, w_{N_s}, \gamma, \delta\}$ can be tuned so as to optimize an evaluation metric on a development set.

Decoding with an n-gram language model requires expanding the lattice to distinguish paths with unique n-gram contexts before LM scores can be assigned the arcs. Using long n-gram context may require pruning to reduce memory usage. Given uniform initial system weights, pruning may remove desirable paths. In this work, the lattices were expanded to bi-gram context and no pruning was performed. A set of bi-gram decoding weights were tuned directly on the expanded lattices using a distributed optimizer (Rosti et al., 2010). Since the score in Equation 2 is not a simple log-linear interpolation, the standard minimum error rate training (Och, 2003) with exact line search cannot be used. Instead, downhill simplex (Press et al., 2007) was used in the optimizer client. After bi-gram decoding weight optimization, another set of 5-gram rescoring weights were tuned on 300-best lists generated from the bi-gram expanded lattices.

3 Hypothesis Alignment Algorithms

Two different methods have been proposed for building confusion networks: pairwise and incremental alignment. In pairwise alignment, each hypothesis corresponding to a source sentence is aligned independently with the skeleton hypothesis. This set of alignments is consolidated using the skeleton words as anchors to form the confusion network (Matusov et al., 2006; Sim et al., 2007). The same word in two hypotheses may be aligned with a different word in the skeleton resulting in repetition in the network. A two-pass alignment algorithm to improve pairwise TER alignments was introduced in (Ayan et al., 2008). In incremental alignment (Rosti et al., 2008), the confusion network is initialized by forming a simple graph with one word per link from the skeleton hypothesis. Each remaining hypothesis is aligned with the partial confusion network, which allows words from all previous hypotheses be considered as matches. The order in which the hypotheses are aligned may influence the alignment quality. Rosti et al. (2009) proposed a sentence specific alignment order by choosing the unaligned hypothesis closest to the partial confusion network according to TER. The following five alignment algorithms were used in this study.

[2]A link is used as a synonym to the set of arcs between two consecutive nodes. The name refers to the confusion network structure's resemblance to a sausage.

3.1 Pairwise GIZA++ Enhanced Hypothesis Alignment

Matusov et al. (2006) proposed using the GIZA++ Toolkit (Och and Ney, 2003) to align a set of target language translations. A parallel corpus where each system output acting as a skeleton appears as a translation of all system outputs corresponding to the same source sentence. The IBM Model 1 (Brown et al., 1993) and hidden Markov model (HMM) (Vogel et al., 1996) are used to estimate the alignment. Alignments from both "translation" directions are used to obtain symmetrized alignments by interpolating the HMM occupation statistics (Matusov et al., 2004). The algorithm may benefit from the fact that it considers the entire test set when estimating the alignment model parameters; i.e., word alignment links from all output sentences influence the estimation, whereas other alignment algorithms only consider words within a pair of sentences (pairwise alignment) or all outputs corresponding to a single source sentence (incremental alignment). However, it does not naturally extend to incremental alignment. The monotone one-to-one alignments are then transformed into a confusion network. This aligner is referred to as GIZA later in this paper.

3.2 Incremental Indirect Hidden Markov Model Alignment

He et al. (2008) proposed using an indirect hidden Markov model (IHMM) for pairwise alignment of system outputs. The parameters of the IHMM are estimated indirectly from a variety of sources including semantic word similarity, surface word similarity, and a distance-based distortion penalty. The alignment between two target language outputs are treated as the hidden states. A standard Viterbi algorithm is used to infer the alignment. The pairwise IHMM was extended to operate incrementally in (Li et al., 2009). Sentence specific alignment order is not used by this aligner, which is referred to as iIHMM later in this paper.

3.3 Incremental Inversion Transduction Grammar Alignment with Flexible Matching

Karakos et al. (2008) proposed using inversion transduction grammars (ITG) (Wu, 1997) for pairwise alignment of system outputs. ITGs form an edit distance, invWER (Leusch et al., 2003), that permits properly nested block movements of substrings. For well-formed sentences, this may be more natural than allowing arbitrary shifts. The ITG algorithm is very expensive due to its $O(n^6)$ complexity. The search algorithm for the best ITG alignment, a best-first chart parsing (Charniak et al., 1998), was augmented with an A^* search heuristic of quadratic complexity (Klein and Manning, 2003), resulting in significant reduction in computational complexity. The finite state-machine heuristic computes a lower bound to the alignment cost of two strings by allowing arbitrary word re-orderings. The ITG hypothesis alignment algorithm was extended to operate incrementally in (Karakos et al., 2010) and a novel version where the cost function is computed based on the stem/synonym similarity of (Snover et al., 2009) was used in this work. Also, a sentence specific alignment order was used. This aligner is referred to as iITGp later in this paper.

3.4 Incremental Translation Edit Rate Alignment with Flexible Matching

Sim et al. (2007) proposed using translation edit rate scorer[3] to obtain pairwise alignment of system outputs. The TER scorer tries to find shifts of blocks of words that minimize the edit distance between the shifted reference and a hypothesis. Due to the computational complexity, a set of heuristics is used to reduce the run time (Snover et al., 2006). The pairwise TER hypothesis alignment algorithm was extended to operate incrementally in (Rosti et al., 2008) and also extended to consider synonym and stem matches in (Rosti et al., 2009). The shift heuristics were relaxed for flexible matching to allow shifts of blocks of words as long as the edit distance is decreased even if there is no exact match in the new position. A sentence specific alignment order was used by this aligner, which is referred to as iTER later in this paper.

3.5 Incremental Translation Edit Rate Plus Alignment

Snover et al. (2009) extended TER scoring to consider synonyms and paraphrase matches, called

[3] http://www.cs.umd.edu/~snover/tercom/

TER-plus (TERp). The shift heuristics in TERp were also relaxed relative to TER. Shifts are allowed if the words being shifted are: (i) exactly the same, (ii) synonyms, stems or paraphrases of the corresponding reference words, or (iii) any such combination. Xu et al. (2011) proposed using an incremental version of TERp for building consensus networks. A sentence specific alignment order was used by this aligner, which is referred to as iTERp later in this paper.

4 Experimental Evaluation

Combination experiments were performed on (i) Arabic-English, from the informal system combination track of the 2009 NIST Open MT Evaluation[4]; (ii) German-English from the system combination evaluation of the 2011 Workshop on Statistical Machine Translation (Callison-Burch et al., 2011) (WMT11) and (iii) Spanish-English, again from WMT11. Eight top-performing systems (as evaluated using case-insensitive BLEU) were used in each language pair. Case insensitive BLEU scores for the individual system outputs on the tuning and test sets are shown in Table 1. About 300 and 800 sentences with four reference translations were available for Arabic-English tune and test sets, respectively, and about 500 and 2500 sentences with a single reference translation were available for both German-English and Spanish-English tune and test sets. The system outputs were lower-cased and tokenized before building confusion networks using the five hypothesis alignment algorithms described above. Unpruned English bi-gram and 5-gram language models were trained with about 6 billion words available for these evaluations. Multiple component language models were trained after dividing the monolingual corpora by source. Separate sets of interpolation weights were tuned for the NIST and WMT experiments to minimize perplexity on the English reference translations of the previous evaluations, NIST MT08 and WMT10. The system combination weights, both bi-gram lattice decoding and 5-gram 300-best list re-scoring weights, were tuned separately for lattices build with each hypothesis alignment algorithm. The final re-scoring outputs were detokenized before computing case insensitive BLEU scores. Statistical significance was computed for each pairwise comparison using bootstrapping (Koehn, 2004).

[4] http://www.itl.nist.gov/iad/mig/tests/mt/2009/ResultsRelease/indexISC.html

	Decode		Oracle	
Aligner	tune	test	tune	test
GIZA	60.06	57.95	75.06	74.47
iTER	59.74	58.63†	73.84	73.20
iTERp	60.18	59.05†	76.43	75.58
iIHMM	60.51	59.27†‡	76.50	76.17
iITGp	60.65	59.37†‡	76.53	76.05

Table 2: Case insensitive BLEU scores for NIST MT09 Arabic-English system combination outputs. Note, four reference translations were available. Decode corresponds to results after weight tuning and Oracle corresponds to graph TER oracle. Dagger (†) denotes statistically significant difference compared to GIZA and double dagger (‡) compared to iTERp and the aligners above it.

The BLEU scores for Arabic-English system combination outputs are shown in Table 2. The first column (Decode) shows the scores on tune and test sets for the decoding outputs. The second column (Oracle) shows the scores for oracle hypotheses obtained by aligning the reference translations with the confusion networks and choosing the path with lowest graph TER (Rosti et al., 2008). The rows representing different aligners are sorted according to the test set decoding scores. The order of the BLEU scores for the oracle translations do not always follow the order for the decoding outputs. This may be due to differences in the compactness of the confusion networks. A more compact network has fewer paths and is therefore less likely to contain significant parts of the reference translation, whereas a reference translation may be generated from a less compact network. On Arabic-English, all incremental alignment algorithms are significantly better than the pairwise GIZA, incremental IHMM and ITG with flexible matching are significantly better than all other algorithms, but not significantly different from each other. The incremental TER and TERp were statistically indistinguishable. Without flexible matching, iITG yields a BLEU score of 58.85 on test. The absolute BLEU gain over the best individual system was between 6.2 and 7.6 points on the test set.

	Arabic		German		Spanish	
System	tune	test	tune	test	tune	test
A	48.84	48.54	21.96	21.41	27.71	27.13
B	49.15	48.97	22.61	21.80	28.42	27.90
C	49.30	49.50	22.77	21.99	28.57	28.23
D	49.38	49.59	22.90	22.41	29.00	28.41
E	49.42	49.75	22.90	22.65	29.15	28.50
F	50.28	50.69	22.98	22.65	29.53	28.61
G	51.49	50.81	23.41	23.06	29.89	29.82
H	51.72	51.74	24.28	24.16	30.55	30.14

Table 1: Case insensitive BLEU scores for the individual system outputs on the tune and test sets for all three source languages.

	Decode		Oracle	
Aligner	tune	test	tune	test
GIZA	25.93	26.02	37.32	38.22
iTERp	26.46	26.10	38.16	38.76
iTER	26.27	26.39†	37.00	37.66
iIHMM	26.34	26.40†	37.87	38.48
iITGp	26.47	26.50†	37.99	38.60

Table 3: Case insensitive BLEU scores for WMT11 German-English system combination outputs. Note, only a single reference translation per segment was available. Decode corresponds to results after weight tuning and Oracle corresponds to graph TER oracle. Dagger (†) denotes statistically significant difference compared to iTERp and GIZA.

	Decode		Oracle	
Aligner	tune	test	tune	test
iTERp	34.20	33.61	50.45	51.28
GIZA	34.02	33.62	50.23	51.20
iTER	34.44	33.79	50.39	50.39
iITGp	34.41	33.85	50.55	51.33
iIHMM	34.61	34.05†	50.48	51.27

Table 4: Case insensitive BLEU scores for WMT11 Spanish-English system combination outputs. Note, only a single reference translation per segment was available. Decode corresponds to results after weight tuning and Oracle corresponds to graph TER oracle. Dagger (†) denotes statistically significant difference compared to aligners above iIHMM.

The BLEU scores for German-English system combination outputs are shown in Table 3. Again, the graph TER oracle scores do not follow the same order as the decoding scores. The scores for GIZA and iTERp are statistically indistinguishable, and iTER, iIHMM, and iITGp are significantly better than the first two. However, they are not statistically different from each other. Without flexible matching, iITG yields a BLEU score of 26.47 on test. The absolute BLEU gain over the best individual system was between 1.9 and 2.3 points on the test set.

The BLEU scores for Spanish-English system combination outputs are shown in Table 4. All aligners but iIHMM are statistically indistinguishable and iIHMM is significantly better than all other aligners. Without flexible matching, iITG yields a BLEU score of 33.62 on test. The absolute BLEU gain over the best individual system was between 3.5 and 3.9 points on the test set.

5 Error Analysis

Error analysis was performed to better understand the gains from system combination. Specifically, (i) how the different types of translation errors are affected by system combination was investigated; and (ii) an attempt to quantify the correlation between the word agreement that results from the different aligners and the translation error, as measured by TER (Snover et al., 2006), was made.

5.1 Influence on Error Types

For each one of the individual systems, and for each one of the three language pairs, the per-sentence errors that resulted from that system, as well as from each one of the the different aligners studied in this paper, were computed. The errors were broken

down into insertions/deletions/substitutions/shifts based on the TER scorer.

The error counts at the document level were aggregated. For each document in each collection, the number of errors of each type that resulted from each individual system as well as each system combination were measured, and their difference was computed. If the differences are mostly positive, then it can be said (with some confidence) that system combination has a significant impact in reducing the error of that type. A paired Wilcoxon test was performed and the p-value that quantifies the probability that the measured error reduction was achieved under the null hypothesis that the system combination performs as well as the best system was computed.

Table 5 shows all conditions under consideration. All cases where the p-value is below 10^{-2} are considered statistically significant. Two observations are in order: (i) all alignment schemes significantly reduce the number of substitution/shift errors; (ii) in the case of insertions/deletions, there is no clear trend; there are cases where the system combination increases the number of insertions/deletions, compared to the individual systems.

5.2 Relationship between Word Agreement and Translation Error

This set of experiments aimed to quantify the relationship between the translation error rate and the amount of agreement that resulted from each alignment scheme. The amount of system agreement at a level x is measured by the number of cases (confusion network arcs) where x system outputs contribute the same word in a confusion network bin. For example, the agreement at level 2 is equal to 2 in Figure 1 because there are exactly 2 arcs (with words "twelve" and "blue") that resulted from the agreement of 2 systems. Similarly, the agreement at level 3 is 1, because there is only 1 arc (with word "cars") that resulted from the agreement of 3 systems. It is hypothesized that a sufficiently high level of agreement should be indicative of the correctness of a word (and thus indicative of lower TER). The agreement statistics were grouped into two values: the "weak" agreement statistic, where at most half of the combined systems contribute a word, and the "strong" agreement statistic, where more than half of the combined systems contribute a word. To signify the fact that real words and "NULL" tokens have different roles and should be treated separately, two sets of agreement statistics were computed.

	non-NULL words		NULL words	
	weak	strong	weak	strong
Arabic	0.087	-0.068	0.192	0.094
German	0.117	-0.067	0.206	0.147
Spanish	0.085	-0.134	0.323	0.102

Table 6: Regression coefficients of the "strong" and "weak" agreement features, as computed with a generalized linear model, using TER as the target variable.

A regression with a generalized linear model (glm) that computed the coefficients of the agreement quantities (as explained above) for each alignment scheme, using TER as the target variable, was performed. Table 6 shows the regression coefficients; they are all significant at p-value < 0.001. As is clear from this table, the negative coefficient of the "strong" agreement quantity for the non-NULL words points to the fact that good aligners tend to result in reductions in translation error. Furthermore, increasing agreements on NULL tokens does not seem to reduce TER.

6 Conclusions

This paper presented a systematic comparison of five different hypothesis alignment algorithms for MT system combination via confusion network decoding. Pre-processing, decoding, and weight tuning were controlled and only the alignment algorithm was varied. Translation quality was compared qualitatively using case insensitive BLEU scores. The results showed that confusion network decoding yields a significant gain over the best individual system irrespective of the alignment algorithm. Differences between the combination output using different alignment algorithms were relatively small, but incremental alignment consistently yielded better translation quality compared to pairwise alignment based on these experiments and previously published literature. Incremental IHMM and a novel incremental ITG with flexible matching consistently yield highest quality combination outputs. Furthermore, an error analysis shows that most of the per-

Language	Aligner	ins	del	sub	shft
Arabic	GIZA	2.2e-16	0.9999	2.2e-16	2.2e-16
	iHMM	2.2e-16	0.433	2.2e-16	2.2e-16
	iITGp	0.8279	2.2e-16	2.2e-16	2.2e-16
	iTER	4.994e-07	3.424e-11	2.2e-16	2.2e-16
	iTERp	2.2e-16	1	2.2e-16	2.2e-16
German	GIZA	7.017e-12	2.588e-06	2.2e-16	2.2e-16
	iHMM	6.858e-07	0.4208	2.2e-16	2.2e-16
	iITGp	0.8551	0.2848	2.2e-16	2.2e-16
	iTER	0.2491	1.233e-07	2.2e-16	2.2e-16
	iTERp	0.9997	0.007489	2.2e-16	2.2e-16
Spanish	GIZA	2.2e-16	0.8804	2.2e-16	2.2e-16
	iHMM	2.2e-16	1	2.2e-16	2.2e-16
	iITGp	2.2e-16	0.9999	2.2e-16	2.2e-16
	iTER	2.2e-16	1	2.2e-16	2.2e-16
	iTERp	3.335e-16	1	2.2e-16	2.2e-16

Table 5: p-values which show which error types are statistically significantly improved for each language and aligner.

formance gains from system combination can be attributed to reductions in substitution errors and word re-ordering errors. Finally, better alignments of system outputs, which tend to cause higher agreement rates on words, correlate with reductions in translation error.

References

Necip Fazil Ayan, Jing Zheng, and Wen Wang. 2008. Improving alignments for better confusion networks for combining machine translation systems. In *Proc. Coling*, pages 33–40.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proc. ASRU*, pages 351–354.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proc. WMT*, pages 22–64.

Eugene Charniak, Sharon Goldwater, and Mark Johnson. 1998. Edge-based best-first chart parsing. In *Proc. Sixth Workshop on Very Large Corpora*, pages 127–133. Morgan Kaufmann.

Jacob Devlin, Antti-Veikko I. Rosti, Shankar Ananthakrishnan, and Spyros Matsoukas. 2011. System combination using discriminative cross-adaptation. In *Proc. IJCNLP*, pages 667–675.

Jonathan G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proc. ASRU*, pages 347–354.

Robert Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *Proc. ANLP*, pages 95–100.

Xiaodong He and Kristina Toutanova. 2009. Joint optimization for machine translation system combination. In *Proc. EMNLP*, pages 1202–1211.

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 *Proc. EMNLP*, pages 98–107.

Almut S. Hildebrand and Stephan Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *AMTA*, pages 254–261.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proc. EAMT*.

Damianos Karakos, Jason Eisner, Sanjeev Khudanpur, and Markus Dreyer. 2008. Machine translation system combination using ITG-based alignments. In *Proc. ACL*, pages 81–84.

Damianos Karakos, Jason R. Smith, and Sanjeev Khudanpur. 2010. Hypothesis ranking and two-pass approaches for machine translation system combination. In *Proc. ICASSP*.

Dan Klein and Christopher D. Manning. 2003. A* parsing: Fast exact Viterbi parse selection. In *Proc. NAACL*, pages 40–47.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proc. MT Summit 2003*, pages 240–247, September.

Chi-Ho Li, Xiaodong He, Yupeng Liu, and Ning Xi. 2009. Incremental hmm alignment for mt system combination. In *Proc. ACL/IJCNLP*, pages 949–957.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400.

Evgeny Matusov, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical machine translation. In *Proc. COLING*, pages 219–225.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proc. EACL*, pages 33–40.

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

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2007. *Numerical recipes: the art of scientific computing*. Cambridge University Press, 3rd edition.

Antti-Veikko I. Rosti, Spyros Matsoukas, and Rirchard Schwartz. 2007a. Improved word-level system combination for machine translation. In *Proc. ACL*, pages 312–319.

Antti-Veikko I. Rosti, Bing Xiang, Spyros Matsoukas, Richard Schwartz, Necip Fazil Ayan, and Bonnie J. Dorr. 2007b. Combining outputs from multiple machine translation systems. In *Proc. NAACL-HLT*, pages 228–235.

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.

Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2009. Incremental hypothesis alignment with flexible matching for building confusion networks: BBN system description for WMT09 system combination task. In *Proc. WMT*, pages 61–65.

Antti-Veikko I. Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. BBN system description for WMT10 system combination task. In *Proc. WMT*, pages 321–326.

Antti-Veikko I. Rosti, Evgeny Matusov, Jason Smith, Necip Fazil Ayan, Jason Eisner, Damianos Karakos, Sanjeev Khudanpur, Gregor Leusch, Zhifei Li, Spyros Matsoukas, Hermann Ney, Richard Schwartz, Bing Zhang, and Jing Zheng. 2011. Confusion network decoding for MT system combination. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 333–361. Springer.

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 *Proc. ICASSP*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciula, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. AMTA*, 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 *Proc. WMT*, pages 259–268.

Stephan Vogel, Hermann Ney, and Christoph Tillman. 1996. HMM-based word alignment in statistical translation. In *Proc. ICCL*, pages 836–841.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, September.

Daguang Xu, Yuan Cao, and Damianos Karakos. 2011. Description of the JHU system combination scheme for WMT 2011. In *Proc. WMT*, pages 171–176.