# Machine Translation System Combination by Confusion Forest

**Taro Watanabe** and **Eiichiro Sumita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Keihanna Science City, 619-0289 JAPAN
{taro.watanabe,eiichiro.sumita}@nict.go.jp

## Abstract

The state-of-the-art system combination method for machine translation (MT) is based on confusion networks constructed by aligning hypotheses with regard to word similarities. We introduce a novel system combination framework in which hypotheses are encoded as a confusion forest, a packed forest representing alternative trees. The forest is generated using syntactic consensus among parsed hypotheses: First, MT outputs are parsed. Second, a context free grammar is learned by extracting a set of rules that constitute the parse trees. Third, a packed forest is generated starting from the root symbol of the extracted grammar through non-terminal rewriting. The new hypothesis is produced by searching the best derivation in the forest. Experimental results on the WMT10 system combination shared task yield comparable performance to the conventional confusion network based method with smaller space.

## 1 Introduction

System combination techniques take the advantages of consensus among multiple systems and have been widely used in fields, such as speech recognition (Fiscus, 1997; Mangu et al., 2000) or parsing (Henderson and Brill, 1999). One of the state-of-the-art system combination methods for MT is based on confusion networks, which are compact graph-based structures representing multiple hypotheses (Bangalore et al., 2001).

Confusion networks are constructed based on string similarity information. First, one skeleton or backbone sentence is selected. Then, other hypotheses are aligned against the skeleton, forming a lattice with each arc representing alternative word candidates. The alignment method is either model-based (Matusov et al., 2006; He et al., 2008) in which a statistical word aligner is used to compute hypothesis alignment, or edit-based (Jayaraman and Lavie, 2005; Sim et al., 2007) in which alignment is measured by an evaluation metric, such as translation error rate (TER) (Snover et al., 2006). The new translation hypothesis is generated by selecting the best path through the network.

We present a novel method for system combination which exploits the syntactic similarity of system outputs. Instead of constructing a string-based confusion network, we generate a packed forest (Billot and Lang, 1989; Mi et al., 2008) which encodes exponentially many parse trees in a polynomial space. The packed forest, or *confusion forest*, is constructed by merging the MT outputs with regard to their syntactic consensus. We employ a grammar-based method to generate the confusion forest: First, system outputs are parsed. Second, a set of rules are extracted from the parse trees. Third, a packed forest is generated using a variant of Earley's algorithm (Earley, 1970) starting from the unique root symbol. New hypotheses are selected by searching the best derivation in the forest. The grammar, a set of rules, is limited to those found in the parse trees. Spurious ambiguity during the generation step is further reduced by encoding the tree local contextual information in each non-terminal symbol, such as parent and sibling labels, using the state representation in Earley's algorithm.



*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1249–1257,
Portland, Oregon, June 19-24, 2011. 

Experiments were carried out for the system combination task of the fifth workshop on statistical machine translation (WMT10) in four directions, {Czech, French, German, Spanish}-to-English (Callison-Burch et al., 2010), and we found comparable performance to the conventional confusion network based system combination in two language pairs, and statistically significant improvements in the others.

First, we will review the state-of-the-art method which is a system combination framework based on confusion networks (§2). Then, we will introduce a novel system combination method based on confusion forest (§3) and present related work in consensus translations (§4). Experiments are presented in Section 5 followed by discussion and our conclusion.

## 2 Combination by Confusion Network

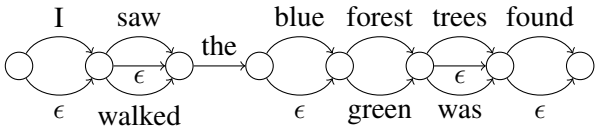
The system combination framework based on confusion network starts from computing pairwise alignment between hypotheses by taking one hypothesis as a reference. Matusov et al. (2006) employs a model based approach in which a statistical word aligner, such as GIZA++ (Och and Ney, 2003), is used to align the hypotheses. Sim et al. (2007) introduced TER (Snover et al., 2006) to measure the edit-based alignment.

Then, one hypothesis is selected, for example by employing a minimum Bayes risk criterion (Sim et al., 2007), as a skeleton, or a backbone, which serves as a building block for aligning the rest of the hypotheses. Other hypotheses are aligned against the skeleton using the pairwise alignment. Figure 1(b) illustrates an example of a confusion network constructed from the four hypotheses in Figure 1(a), assuming the first hypothesis is selected as our skeleton. The network consists of several arcs, each of which represents an alternative word at that position, including the empty symbol, $\epsilon$.

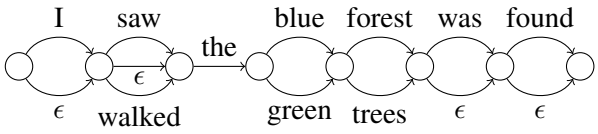
This pairwise alignment strategy is prone to spurious insertions and repetitions due to alignment errors such as in Figure 1(a) in which "green" in the third hypothesis is aligned with "forest" in the skeleton. Rosti et al. (2008) introduces an incremental method so that hypotheses are aligned incrementally to the growing confusion network, not only the skeleton hypothesis. In our example, "green trees" is aligned with "blue forest" in Figure 1(c).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| * | I | saw | the | | forest | | |
| | I | walked | the | blue | forest | | |
| | I | saw | the | | green | trees | |
| | | | the | | forest | was | found |

(a) Pairwise alignment using the first starred hypothesis as a skeleton.




(b) Confusion network from (a)




(c) Incrementally constructed confusion network

Figure 1: An example confusion network construction

The confusion network construction is largely influenced by the skeleton selection, which determines the global word reordering of a new hypothesis. For example, the last hypothesis in Figure 1(a) has a passive voice grammatical construction while the others are active voice. This large grammatical difference may produce a longer sentence with spuriously inserted words, as in "I saw the blue trees was found" in Figure 1(c). Rosti et al. (2007b) partially resolved the problem by constructing a large network in which each hypothesis was treated as a skeleton and the multiple networks were merged into a single network.

## 3 Combination by Confusion Forest

The confusion network approach to system combination encodes multiple hypotheses into a compact lattice structure by using word-level consensus. Likewise, we propose to encode multiple hypotheses into a confusion forest, which is a packed forest which represents multiple parse trees in a polynomial space (Billot and Lang, 1989; Mi et al., 2008) Syntactic consensus is realized by sharing tree frag-

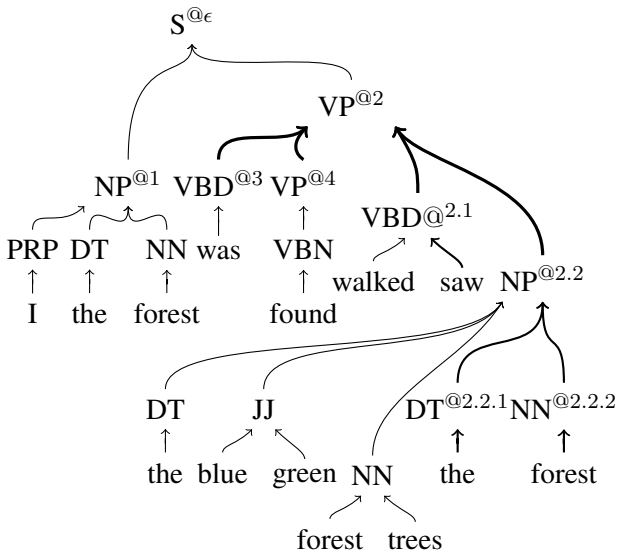



Figure 2: An example packed forest representing hypotheses in Figure 1(a).

ments among parse trees. The forest is represented as a hypergraph which is exploited in parsing (Klein and Manning, 2001; Huang and Chiang, 2005) and machine translation (Chiang, 2007; Huang and Chiang, 2007).

More formally, a hypergraph is a pair $\langle V, E\rangle$ where $V$ is the set of nodes and $E$ is the set of hyperedges. Each node in $V$ is represented as $X^{@p}$ where $\mathrm{X} \in \mathcal{N}$ is a non-terminal symbol and $p$ is an address (Shieber et al., 1995) that encapsulates each node id relative to its parent. The root node is given the address $\epsilon$ and the address of the first child of node $p$ is given $p.1$. Each hyperedge $e \in E$ is represented as a pair $\langle head(e), tails(e)\rangle$ where $head(e) \in V$ is a head node and $tails(e) \in V^*$ is a list of tail nodes, corresponding to the left-hand side and the right-hand side of an instance of a rule in a CFG, respectively. Figure 2 presents an example packed forest for the parsed hypotheses in Figure 1(a). For example, $\mathrm{VP}^{@2}$ has two hyperedges, $\langle \mathrm{VP}^{@2}, (\mathrm{VBD}^{@3}, \mathrm{VP}^{@4})\rangle$ and $\langle \mathrm{VP}^{@2}, (\mathrm{VBD}^{@2.1}, \mathrm{NP}^{@2.2})\rangle$, leading to different derivations where the former takes the grammatical construction in passive voice while the latter in active voice.

Given system outputs, we employ the following grammar based approach for constructing a confusion forest: First, MT outputs are parsed. Second, a grammar is learned by treating each hyperedge as an instance of a CFG rule. Third, a forest is generated from the unique root symbol of the extracted grammar through non-terminal rewriting.

Initialization:

$$\frac{}{[\mathrm{TOP} \rightarrow \bullet \mathrm{S}, 0] : \bar{1}}$$

Scan:

$$\frac{[\mathrm{X} \rightarrow \alpha \bullet x\beta, h] : u}{[\mathrm{X} \rightarrow \alpha x \bullet \beta, h] : u}$$

Predict:

$$\frac{[\mathrm{X} \rightarrow \alpha \bullet \mathrm{Y}\beta, h]}{[\mathrm{Y} \rightarrow \bullet\gamma, h+1] : u} \quad \mathrm{Y} \xrightarrow{u} \gamma \in \mathcal{G}, h < H$$

Complete:

$$\frac{[\mathrm{X} \rightarrow \alpha \bullet \mathrm{Y}\beta, h] : u \quad [\mathrm{Y} \rightarrow \gamma\bullet, h+1] : v}{[\mathrm{X} \rightarrow \alpha\mathrm{Y} \bullet \beta, h] : u \otimes v}$$

Goal:

$$[\mathrm{TOP} \rightarrow \mathrm{S}\bullet, 0]$$

Figure 3: The deductive system for Earley's generation algorithm

### 3.1 Forest Generation

Given the extracted grammar, we apply a variant of Earley's algorithm (Earley, 1970) which can generate strings in a left-to-right manner from the unique root symbol, TOP. Figure 3 presents the deductive inference rules (Goodman, 1999) for our generation algorithm. We use capital letters $\mathrm{X} \in \mathcal{N}$ to denote non-terminals and $x \in \mathcal{T}$ for terminals. Lowercase Greek letters $\alpha$, $\beta$ and $\gamma$ are strings of terminals and non-terminals $(\mathcal{T} \cup \mathcal{N})^*$. $u$ and $v$ are weights associated with each item.

The major difference compared to Earley's parsing algorithm is that we ignore the terminal span information each non-terminal covers and keep track of the height of derivations by $h$. The scanning step will always succeed by moving the dot to the right. Combined with the prediction and completion steps, our algorithm may potentially generate a spuriously deep forest. Thus, the height of the forest is constrained in the prediction step not to exceed $H$, which is empirically set to 1.5 times the maximum

height of the parsed system outputs.

### 3.2 Tree Annotation

The grammar compiled from the parsed trees is local in that it can represent a finite number of sentences translated from a specific input sentence. Although its coverage is limited, our generation algorithm may yield a spuriously large forest. As a way to reduce spurious ambiguities, we relabel the non-terminal symbols assigned to each parse tree before extracting rules.

Here, we replace each non-terminal symbol by the state representation of Earley's algorithm corresponding to the sequence of prediction steps starting from TOP. Figure 4(a) presents an example parse tree with each symbol replaced by the Earley's state in Figure 4(b). For example, the label for VBD is replaced by •S + NP : •VP + •VBD : NP which corresponds to the prediction steps of TOP → •S, S → NP • VP and VP → •VBD NP. The context represented in the Earley's state is further limited by the vertical and horizontal Markovization (Klein and Manning, 2003). We define the vertical order $v$ in which the label is limited to memorize only $v$ previous prediction steps. For instance, setting $v = 1$ yields NP : •VP + •VBD : NP in our example. Likewise, we introduce the horizontal order $h$ which limits the number of sibling labels memorized on the left and the right of the dotted label. Limiting $h = 1$ implies that each deductive step is encoded with at most three symbols.

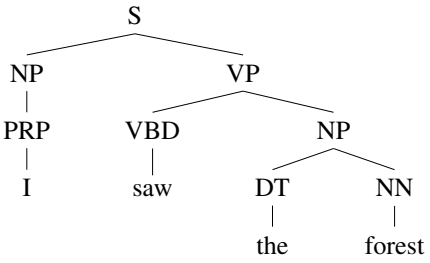
No limits in the horizontal and vertical Markovization orders implies memorizing of all the deductions and yields a confusion forest representing the union of parse trees through the grammar collection and the generation processes. More relaxed horizontal orders allow more reordering of subtrees in a confusion forest by discarding the sibling context in each prediction step. Likewise, constraining the vertical order generates a deeper forest by ignoring the sequence of symbols leading to a particular node.

### 3.3 Forest Rescoring

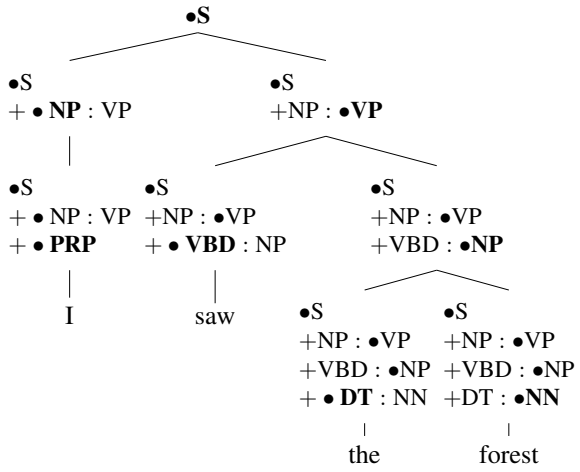
From the packed forest $F$, new $k$-best derivations are extracted from all possible derivations $D$ by efficient forest-based algorithms for $k$-best parsing (Huang and Chiang, 2005). We use a linear combination of features as our objective function to seek for the best derivation $\hat{d}$:

$$\hat{d} = \arg\max_{d \in D} \mathbf{w}^\top \cdot \mathbf{h}(d, F) \quad (1)$$

where $\mathbf{h}(d, F)$ is a set of feature functions scaled by weight vector $\mathbf{w}$. We use cube-pruning (Chiang, 2007; Huang and Chiang, 2007) to approximately intersect with non-local features, such as $n$-gram language models. Then, $k$-best derivations are extracted from the rescored forest using algorithm 3 of Huang and Chiang (2005).




(a) A parse tree for "I saw the forest"




(b) Earley's state annotated tree for (a). The sub-labels in boldface indicate the original labels.

Figure 4: Label annotation by Earley's alsogirhtm state

## 4 Related Work

Consensus translations have been extensively studied with many granularities. One of the simplest forms is a sentence-based combination in which hypotheses are simply reranked without merging (Nomoto, 2004). Frederking and Nirenburg (1994)

proposed a phrasal combination by merging hypotheses in a chart structure, while others depended on confusion networks, or similar structures, as a building block for merging hypotheses at the word level (Bangalore et al., 2001; Matusov et al., 2006; He et al., 2008; Jayaraman and Lavie, 2005; Sim et al., 2007). Our work is the first to explicitly exploit syntactic similarity for system combination by merging hypotheses into a syntactic packed forest. The confusion forest approach may suffer from parsing errors such as the confusion network construction influenced by alignment errors. Even with parsing errors, we can still take a tree fragment-level consensus as long as a parser is consistent in that similar syntactic mistakes would be made for similar hypotheses.

Rosti et al. (2007a) describe a re-generation approach to consensus translation in which a phrasal translation table is constructed from the MT outputs aligned with an input source sentence. New translations are generated by decoding the source sentence again using the newly extracted phrase table. Our grammar-based approach can be regarded as a re-generation approach in which an off-the-shelf monolingual parser, instead of a word aligner, is used to annotate syntactic information to each hypothesis, then, a new translation is generated from the merged forest, not from the input source sentence through decoding. In terms of generation, our approach is an instance of statistical generation (Langkilde and Knight, 1998; Langkilde, 2000). Instead of generating forests from semantic representations (Langkilde, 2000), we generate forests from a CFG encoding the consensus among parsed hypotheses.

Liu et al. (2009) present joint decoding in which a translation forest is constructed from two distinct MT systems, tree-to-string and string-to-string, by merging forest outputs. Their merging method is either translation-level in which no new translation is generated, or derivation-level in that the rules sharing the same left-hand-side are used in both systems. While our work is similar in that a new forest is constructed by sharing rules among systems, although their work involves no consensus translation and requires structures internal to each system such as model combinations (DeNero et al., 2010).

| | | cz-en | de-en | es-en | fr-en |
|---|---|---|---|---|---|
| # of systems | | 6 | 16 | 8 | 14 |
| avg. words | tune | 10.6K | 10.9K | 10.9K | 11.0K |
| | test | 50.5K | 52.1K | 52.1K | 52.4K |
| sentences | tune | 455 | | | |
| | test | 2,034 | | | |

Table 1: WMT10 system combination tuning/testing data

## 5 Experiments

### 5.1 Setup

We ran our experiments for the WMT10 system combination task usinge four language pairs, {Czech, French, German, Spanish}-to-English (Callison-Burch et al., 2010). The data is summarized in Table 1. The system outputs are retokenized to match the Penn-treebank standard, parsed by the Stanford Parser (Klein and Manning, 2003), and lower-cased.

We implemented our confusion forest system combination using an in-house developed hypergraph-based toolkit *cicada* which is motivated by generic weighted logic programming (Lopez, 2009), originally developed for a synchronous-CFG based machine translation system (Chiang, 2007). Input to our system is a collection of hypergraphs, a set of parsed hypotheses, from which rules are extracted and a new forest is generated as described in Section 3. Our baseline, also implemented in *cicada*, is a confusion network-based system combination method (§2) which incrementally aligns hypotheses to the growing network using TER (Rosti et al., 2008) and merges multiple networks into a large single network. After performing epsilon removal, the network is transformed into a forest by parsing with monotone rules of S $\rightarrow$ X, S $\rightarrow$ S X and X $\rightarrow$ $x$. $k$-best translations are extracted from the forest using the forest-based algorithms in Section 3.3.

### 5.2 Features

The feature weight vector $\mathbf{w}$ in Equation 1 is tuned by MERT over hypergraphs (Kumar et al., 2009).

We use three lower-cased 5-gram language mod-

els $h^i_{lm}(d)$: English Gigaword Fourth edition[1], the English side of French-English $10^9$ corpus and the news commentary English data[2]. The count based features $h_t(d)$ and $h_e(d)$ count the number of terminals and the number of hyperedges in $d$, respectively. We employ $M$ confidence measures $h^m_s(d)$ for $M$ systems, which basically count the number of rules used in $d$ originally extracted from $m$th system hypothesis (Rosti et al., 2007a).

Following Macherey and Och (2007), BLEU (Papineni et al., 2002) correlations are also incorporated in our system combination. Given $M$ system outputs $\mathbf{e}_1 ... \mathbf{e}_M$, $M$ BLEU scores are computed for $d$ using each of the system outputs $\mathbf{e}_m$ as a reference

$$h^m_b(d) = BP(\mathbf{e}, \mathbf{e}_m) \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log \rho_n(\mathbf{e}, \mathbf{e}_m)\right)$$

where $\mathbf{e} = \text{yield}(d)$ is a terminal yield of $d$, $BP(\cdot)$ and $\rho_n(\cdot)$ respectively denote brevity penalty and $n$-gram precision. Here, we use approximated unclipped $n$-gram counts (Dreyer et al., 2007) for computing $\rho_n(\cdot)$ with a compact state representation (Li and Khudanpur, 2009).

Our baseline confusion network system has an additional penalty feature, $h_p(m)$, which is the total edits required to construct a confusion network using the $m$th system hypothesis as a skeleton, normalized by the number of nodes in the network (Rosti et al., 2007b).

### 5.3 Results

Table 2 compares our confusion forest approach (CF) with different orders, a confusion network (CN) and max/min systems measured by BLEU (Papineni et al., 2002). We vary the horizontal orders, $h = 1, 2, \infty$ with vertical orders of $v = 3, 4, \infty$. Systems without statistically significant differences from the best result ($p < 0.05$) are indicated by bold face. Setting $v = \infty$ and $h = \infty$ achieves comparable performance to CN. Our best results in three languages come from setting $v = \infty$ and $h = 2$, which favors little reordering of phrasal structures. In general, lower horizontal and vertical order leads to lower BLEU.

[1]LDC catalog No. LDC2009T13

[2]Those data are available from `http://www.statmt.org/wmt10/`.

| language | cz-en | de-en | es-en | fr-en |
|---|---|---|---|---|
| system min | 14.09 | 15.62 | 21.79 | 16.79 |
| max | 23.44 | 24.10 | 29.97 | **29.17** |
| CN | 23.70 | 24.09 | **30.45** | **29.15** |
| CF$_{v=\infty,h=\infty}$ | **24.13** | 24.18 | **30.41** | **29.57** |
| CF$_{v=\infty,h=2}$ | **24.14** | **24.58** | **30.52** | 28.84 |
| CF$_{v=\infty,h=1}$ | **24.01** | 23.91 | **30.46** | **29.32** |
| CF$_{v=4,h=\infty}$ | **23.93** | 23.57 | 29.88 | 28.71 |
| CF$_{v=4,h=2}$ | **23.82** | 22.68 | 29.92 | 28.83 |
| CF$_{v=4,h=1}$ | **23.77** | 21.42 | 30.10 | 28.32 |
| CF$_{v=3,h=\infty}$ | 23.38 | 23.34 | 29.81 | 27.34 |
| CF$_{v=3,h=2}$ | 23.30 | 23.95 | 30.02 | 28.19 |
| CF$_{v=3,h=1}$ | 23.23 | 21.43 | 29.27 | 26.53 |

Table 2: Translation results in lower-case BLEU. CN for confusion network and CF for confusion forest with different vertical ($v$) and horizontal ($h$) Markovization order.

| language | cz-en | de-en | es-en | fr-en |
|---|---|---|---|---|
| rerank | 29.40 | 32.32 | 36.83 | 36.59 |
| CN | 38.52 | 34.97 | 47.65 | 46.37 |
| CF$_{v=\infty,h=\infty}$ | 30.51 | 34.07 | 38.69 | 38.94 |
| CF$_{v=\infty,h=2}$ | 30.61 | 34.25 | 38.87 | 39.10 |
| CF$_{v=\infty,h=1}$ | 31.09 | 34.65 | 39.27 | 39.51 |
| CF$_{v=4,h=\infty}$ | 30.86 | 34.19 | 39.17 | 39.39 |
| CF$_{v=4,h=2}$ | 30.96 | 34.32 | 39.35 | 39.57 |
| CF$_{v=4,h=1}$ | 31.44 | 34.62 | 39.69 | 39.90 |
| CF$_{v=3,h=\infty}$ | 31.03 | 34.30 | 39.29 | 39.57 |
| CF$_{v=3,h=2}$ | 31.25 | 34.97 | 39.61 | 40.00 |
| CF$_{v=3,h=1}$ | 31.55 | 34.60 | 39.72 | 39.97 |

Table 3: Oracle lower-case BLEU

Table 3 presents oracle BLEU achievable by each combination method. The gains achievable by the CF over simple reranking are small, at most 2-3 points, indicating that small variations are encoded in confusion forests. We also observed that a lower horizontal and vertical order leads to better BLEU potentials. As briefly pointed out in Section 3.2, the higher horizontal and vertical order implies more faithfulness to the original parse trees. Introducing new tree fragments to confusion forests leads to new phrasal translations with enlarged forests, as presented in Table 4, measured by the average number

| lang | cz-en | de-en | es-en | fr-en |
|---|---|---|---|---|
| CN | 2,222.68 | 47,231.20 | 2,932.24 | 11,969.40 |
| lattice | 1,723.91 | 41,403.90 | 2,330.04 | 10,119.10 |
| $CF_{v=\infty}$ | 230.08 | 540.03 | 262.30 | 386.79 |
| $CF_{v=4}$ | 254.45 | 651.10 | 302.01 | 477.51 |
| $CF_{v=3}$ | 286.01 | 802.79 | 349.21 | 575.17 |

Table 4: Hypegraph size measured by the average number of hyperedges ($h = 1$ for CF). "lattice" is the average number of edges in the original CN.

of hyperedges[3]. The larger potentials do not imply better translations, probably due to the larger search space with increased search errors. We also conjecture that syntactic variations were not captured by the $n$-gram like string-based features in Section 5.2, therefore resulting in BLEU loss, which will be investigated in future work.

In contrast, CN has more potential for generating better translations, with the exception of the German-to-English direction, with scores that are usually 10 points better than simple sentence-wise reranking. The low potential in German should be interpreted in the light of the extremely large confusion network in Table 4. We postulate that the divergence in German hypotheses yields wrong alignments, and therefore amounts to larger networks with incorrect hypotheses. Table 4 also shows that CN produces a forest that is an order of magnitude larger than those created by CFs. Although we cannot directly relate the runtime and the number of hyperedges in CN and CFs, since the shape of the forests are different, CN requires more space to encode the hypotheses than those by CFs.

Table 5 compares the average length of the minimum/maximum hypothesis that each method can produce. CN may generate shorter hypotheses, whereby CF prefers longer hypotheses as we decrease the vertical order. Large divergence is also observed for German, such as for hypergraph size.

## 6 Conclusion

We presented a confusion forest based method for system combination in which system outputs are merged into a packed forest using their syntactic similarity. The forest construction is treated as a generation from a CFG compiled from the parsed outputs. Our experiments indicate comparable performance to a strong confusion network baseline with smaller space, and statistically significant gains in some language pairs.

To our knowledge, this is the first work to directly introduce syntactic consensus to system combination by encoding multiple system outputs into a single forest structure. We believe that the confusion forest based approach to system combination has future exploration potential. For instance, we did not employ syntactic features in Section 5.2 which would be helpful in discriminating hypotheses in larger forests. We would also like to analyze the trade-offs, if any, between parsing errors and confusion forest constructions by controlling the parsing qualities. As an alternative to the grammar-based forest generation, we are investigating an edit distance measure for tree alignment, such as tree edit distance (Bille, 2005) which basically computes insertion/deletion/replacement of nodes in trees.

[3]We measure the hypergraph size before intersecting with non-local features, like $n$-gram language models.

| language | | cz-en | de-en | es-en | fr-en |
|---|---|---|---|---|---|
| system | avg. | 24.84 | 25.62 | 25.63 | 25.75 |
| CN | min | 11.09 | 3.39 | 12.27 | 7.94 |
| | max | 33.69 | 40.65 | 33.22 | 36.27 |
| $CF_{v=\infty}$ | min | 15.97 | 10.88 | 17.67 | 16.62 |
| | max | 35.20 | 47.20 | 35.28 | 37.94 |
| $CF_{v=4}$ | min | 15.52 | 10.58 | 17.02 | 15.85 |
| | max | 37.11 | 53.67 | 38.56 | 42.64 |
| $CF_{v=3}$ | min | 15.15 | 10.34 | 16.54 | 15.30 |
| | max | 39.88 | 68.45 | 42.85 | 49.55 |

Table 5: Average min/max hypothesis length producible by each method ($h = 1$ for CF).

## Acknowledgments

We would like to thank anonymous reviewers and our colleagues for helpful comments and discussion.

## References

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

Philip Bille. 2005. A survey on tree edit distance and related problems. *Theor. Comput. Sci.*, 337:217–239, June.

Sylvie Billot and Bernard Lang. 1989. The structure of shared forests in ambiguous parsing. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 143–151, Vancouver, British Columbia, Canada, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July. Revised August 2010.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

John DeNero, Shankar Kumar, Ciprian Chelba, and Franz Och. 2010. Model combination for machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 975–983, Los Angeles, California, June.

Markus Dreyer, Keith Hall, and Sanjeev Khudanpur. 2007. Comparing reordering constraints for smt using efficient bleu oracle computation. In *Proceedings of SSST, NAACL-HLT 2007 / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 103–110, Rochester, New York, April.

Jay Earley. 1970. An efficient context-free parsing algorithm. *Communications of the Association for Computing Machinery*, 13:94–102, February.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proceedings of Automatic Speech Recognition and Understanding (ASRU), 1997*, pages 347 –354, December.

Robert Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *Proceedings of the fourth conference on Applied natural language processing*, pages 95–100, Morristown, NJ, USA.

Joshua Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25:573–605, December.

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, Honolulu, Hawaii, October.

John C. Henderson and Eric Brill. 1999. Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the Fourth Conference on Empirical Methods in Natural Language Processing*, pages 187–194.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 53–64, Vancouver, British Columbia, October.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic, June.

Shyamsundar Jayaraman and Alon Lavie. 2005. Multi-engine machine translation guided by explicit word matching. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, ACL '05, pages 101–104, Morristown, NJ, USA.

Dan Klein and Christopher D. Manning. 2001. Parsing and hypergraphs. In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT-2001)*, pages 123–134.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July.

Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 163–171, Suntec, Singapore, August.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL-36, pages 704–710, Morristown, NJ, USA.

Irene Langkilde. 2000. Forest-based statistical sentence generation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 170–177, San Francisco, CA, USA.

Zhifei Li and Sanjeev Khudanpur. 2009. Efficient extraction of oracle-best translations from hypergraphs. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 9–12, Boulder, Colorado, June.

Yang Liu, Haitao Mi, Yang Feng, and Qun Liu. 2009. Joint decoding with multiple translation models. In

*Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 576–584, Suntec, Singapore, August.

Adam Lopez. 2009. Translation as weighted deduction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 532–540, Athens, Greece, March.

Wolfgang Macherey and Franz J. Och. 2007. An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 986–995, Prague, Czech Republic, June.

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

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.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*, pages 192–199, Columbus, Ohio, June.

Tadashi Nomoto. 2004. Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 494–501, Barcelona, Spain, July.

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

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

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007a. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York, April.

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

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.

Stuart M. Shieber, Yves Schabes, and Fernando C. N. Pereira. 1995. Principles and implementation of deductive parsing. *Journal of Logic Programming*, 24(1–2):3–36, July–August.

K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi, and P.C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2007*, volume 4, pages IV–105 –IV–108, April.

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