Jane: Open Source Machine Translation System Combination

Markus Freitag[1] and **Matthias Huck**[2] and **Hermann Ney**[1]

[1] Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{freitag,ney}@cs.rwth-aachen.de

[2] School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK
mhuck@inf.ed.ac.uk

Abstract

Different machine translation engines can be remarkably dissimilar not only with respect to their technical paradigm, but also with respect to the translation output they yield. *System combination* is a method for combining the output of multiple machine translation engines in order to take benefit of the strengths of each of the individual engines.

In this work we introduce a novel system combination implementation which is integrated into Jane, RWTH's open source statistical machine translation toolkit. On the most recent *Workshop on Statistical Machine Translation* system combination shared task, we achieve improvements of up to 0.7 points in BLEU over the best system combination hypotheses which were submitted for the official evaluation. Moreover, we enhance our system combination pipeline with additional *n*-gram language models and lexical translation models.

1 Introduction

We present a novel machine translation system combination framework which has been implemented and released as part of the most recent version of the Jane toolkit.[1] Our system combination framework has already been applied successfully for joining the outputs of different individual machine translation engines from several project partners within large-scale projects like Quaero (Peitz and others, 2013), EU-BRIDGE (Freitag and others, 2013), and DARPA BOLT. The combined translation is typically of better quality than any of the individual hypotheses. The source code of our framework has now been released to the public.

We focus on system combination via confusion network decoding. This basically means that we align all input hypotheses from individual machine translation (MT) engines together and extract a combination as a new output. For our baseline algorithm we only need the first best translation from each of the different MT engines, without any additional information. Supplementary to the baseline models integrated into our framework, we optionally allow for utilization of *n*-gram language models and IBM-1 lexicon models (Brown et al., 1993), both trained on additional training corpora that might be at hand.

We evaluate the Jane system combination framework on the latest official *Workshop on Statistical Machine Translation* (WMT) system combination shared task (Callison-Burch et al., 2011). Many state-of-the-art MT system combination toolkits have been evaluated on this task, which allows us to directly compare the results obtained with our novel Jane system combination framework with the best known results obtained with other toolkits.

The paper is structured as follows: We commence with giving a brief outline of some related work (Section 2). In Section 3 we describe the techniques which are implemented in the Jane MT system combination framework. The experimental results are presented and analyzed in Section 4. We conclude the paper in Section 5.

2 Related Work

The first application of system combination to MT has been presented by Bangalore et al. (2001). They used a multiple string alignment (MSA) approach to align the hypotheses together and built a confusion network from which the system combination output is determined using majority vot-

[1] Jane is publicly available under an open source non-commercial license and can be downloaded from http://www.hltpr.rwth-aachen.de/jane/ .

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 29–32,
Gothenburg, Sweden, April 26-30 2014.

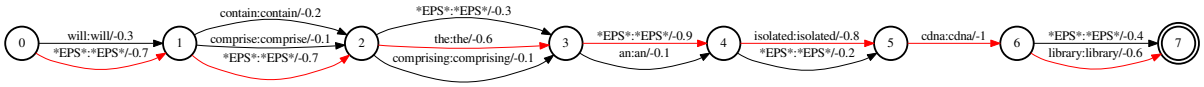

Figure 1: Scored confusion network. *EPS* denotes the empty word, red arcs highlight the shortest path.

ing and an additional language model. Matusov et al. (2006) proposed an alignment based on the GIZA++ toolkit which introduced word reordering not present in MSA, and Sim et al. (2007) used alignments produced by TER scoring (Snover et al., 2006). Extensions of the last two are based on hidden Markov models (He et al., 2008), inversion transduction grammars (Karakos et al., 2008), or METEOR (Heafield and Lavie, 2010).

3 The Jane MT System Combination Framework

In this section we describe the techniques for MT system combination which we implemented in the Jane toolkit.[2] We first address the generation of a confusion network from the input translations. For that we need a pairwise alignment between all input hypotheses. We then present word reordering mechanisms, the baseline models, and additional advanced models which can be applied for system combination using Jane. The system combination decoding step basically involves determining the shortest path through the confusion network based on several model scores from this network.

3.1 Confusion Network

A confusion network represents all different combined translations we can generate from the set of provided input hypotheses. Figure 1 depicts an example of a confusion network. A word alignment between all pairs of input hypotheses is required for generating a confusion network. For convenience, we first select one of the input hypotheses as the primary hypothesis. The primary hypothesis then determines the word order and all remaining hypotheses are word-to-word aligned to the given word order.

To generate a meaningful confusion network, we should adopt an alignment which only allows to switch between words which are synonyms, misspellings, morphological variants or on a higher level paraphrases of the words from the primary hypothesis. In this work we use METEOR alignments. METEOR (Denkowski and Lavie, 2011) was originally designed to reorder a translation for scoring and has a high precision. The recall is lower because synonyms which are not in the METEOR database or punctuation marks like "!" and "?" are not aligned to each other. For our purposes, we augment the METEOR paraphrase table with entries like ".|!", ".|?", or "the|a".

Figure 2 shows an example METEOR hypothesis alignment. The primary hypothesis "isolated cdna lib" determines the word order. An entry "a|b" means that word "a" from a secondary hypothesis has been aligned to word "b" from the primary one. "*EPS*" is the empty word and thus an entry "*EPS*|b" means that no word could be aligned to the primary hypothesis word "b". "a|*EPS*" means that the word "a" has not been aligned to any word from the primary hypothesis.

After producing the alignment information, we can build the confusion network. Now, we are able to not only extract the original primary hypothesis from the confusion network but also switch words from the primary hypothesis to words from any secondary hypothesis (also the empty word) or insert words or sequences of words.

In the final confusion network, we do not stick to one hypothesis as the primary system. For m input hypotheses we build m different confusion networks, each having a different system as primary system. The final confusion network is a union of all m networks.[3]

The most straightforward way to obtain a combined hypothesis from a confusion network is to extract it via majority voting. For example, in the first column in Figure 3, "the" has been seen three times, but the translation options "a" and "an" have each been seen only once. By means of a straight majority vote we would extract "the". As the different single system translations are of varying utility for system combination, we assign a system weight to each input hypothesis. The system weights are set by optimizing scaling factors for binary system voting features (cf. Section 3.3). We employ some more weighted baseline features

[2] Practical usage aspects are explained in the manual: http://www.hltpr.rwth-aachen.de/jane/manual.pdf

[3] Jane's implementation for building confusion networks is based on the OpenFST library (Allauzen et al., 2007).

the|*EPS*	isolated|isolated	cdna|cdna	*EPS*|lib
a|*EPS*	isolated|isolated	cdna|cdna	lib|lib
an|*EPS*	isolated|isolated	cdna|cdna	lib|lib
the|*EPS*	*EPS*|isolated	cdna|cdna	*EPS*|lib
the|*EPS*	*EPS*|isolated	cdna|cdna	lib|lib

Figure 2: Alignment result after running METEOR. *EPS* denotes the empty word.

EPS	isolated	cdna	lib
the	isolated	cdna	*EPS*
a	isolated	cdna	lib
an	isolated	cdna	lib
the	*EPS*	cdna	lib
the	*EPS*	cdna	*EPS*
the	isolated	cdna	lib

Figure 3: Majority vote on aligned words. The last line is the system combination output.

and additional models (cf. Section 3.4) in the decision process. In Figure 1 we scored the confusion network with some system weights and used the shortest path algorithm to find the hypothesis with the highest score (the hypothesis along the path highlighted in red).

3.2 Word Reordering

Many words from secondary hypotheses can be unaligned as they have no connection to any words of the primary hypothesis. However, words from different secondary systems could be related to each other. In order to account for these relations and to give the words from the secondary hypotheses a higher chance to be present in the combined output, we introduce some simple word reordering mechanisms.

We rank the hypotheses according to a language model trained on all input hypotheses. We initialize the confusion network with the sentence from the primary system. During the generation of the confusion network we align the hypotheses consecutively into the confusion network via the following procedure:

- If a word w_i from hypothesis A has a relation to a word v_j of the primary hypothesis, we insert it as a new translation alternative to v_j.
- If w_i has no relation to the primary, but to a word u_k from a secondary hypothesis in the confusion network, we insert w_i as a new translation alternative to u_k.
- Otherwise we insert w_i in front of the previous inserted word w_{i-1} of hypothesis A. The new position gets an epsilon arc for the primary and all unrelated secondary systems.

3.3 Baseline Models

Once we have the final confusion network, we want to adopt models which are valuable features to score the different translation options. In our implementation we use the following set of standard models:

m binary system voting features For each word the voting feature for system i ($1 \leq i \leq m$) is 1 iff the word is from system i, otherwise 0.

Binary primary system feature A feature that marks the primary hypothesis.

LM feature 3-gram language model trained on the input hypotheses.

Word penalty Counts the number of words.

3.4 Additional Models

The `Jane` system combination toolkit also provides the possibility to utilize some additional models for system combination. For the current release we integrated the optional usage of the following additional models:

Big LM A big language model trained on larger monolingual target-side corpora.

IBM-1 Source-to-target and target-to-source IBM-1 lexical translation models obtained from bilingual training data.

4 Experimental Results

All experiments are conducted on the latest official WMT system combination shared task.[4] We exclusively employ resources which were permitted for the constrained track of the task in all our setups. The big LM was trained on News Commentary and Europarl data. As tuning set we used *newssyscombtune2011*, as test set we used *newssyscombtest2011*. Feature weights have been optimized with MERT (Och, 2003). Table 1 contains the empirical results (truecase). For all four language pairs we achieve improvements over the best 2011 evaluation system combination submission either in BLEU or TER. We get the highest improvement of 0.7 points in BLEU for es→en when adding both the big LM and IBM-1 features. Adding the big LM over the baseline enhances the translation quality for all four language pairs. Adding IBM-1 lexicon models on top of the big LM is of marginal or no benefit for most language

[4]The most recent system combination shared task that has been organized as part of the WMT evaluation campaign took place in 2011. `http://www.statmt.org/wmt11/system-combination-task.html`

Table 1: Experimental results on the WMT system combination tasks (newssyscombtest2011).

system	cz→en		de→en		es→en		fr→en	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
best single system	28.7	53.4	23.0	59.5	28.9	51.2	29.4	52.0
best 2011 evaluation syscomb	28.8	55.2	25.1	57.4	32.4	49.9	31.3	50.1
Jane syscomb baseline	28.8	53.6	24.7	57.6	32.7	50.3	31.3	50.3
Jane syscomb + big LM	29.0	54.5	25.0	57.3	32.9	50.3	31.4	50.0
Jane syscomb + big LM + IBM-1	29.0	54.5	25.0	57.3	33.1	50.0	31.5	50.1

pairs, but at least provides slight improvements for es→en.

5 Conclusion

RWTH's open source machine translation toolkit Jane now includes a state-of-the-art system combination framework. We found that the Jane system combination performs on a similar level or better than the best evaluation system combination submissions on all WMT 2011 system combination shared task language pairs (with English as target language). We furthermore presented the effects of integrating a big *n*-gram language model and of lexical features from IBM-1 models.

Acknowledgements

This material is based upon work supported by the DARPA BOLT project under Contract No. HR0011- 12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In Jan Holub and Jan Zdárek, editors, *Implementation and Application of Automata*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer Berlin Heidelberg.

Srinivas Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing Consensus Translation from Multiple Machine Translation Systems. In *Proc. of 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. of WMT*, pages 22–64.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proc. of WMT*, pages 85–91.

Markus Freitag et al. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proc. of IWSLT*.

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. of EMNLP*, pages 98–107.

Kenneth Heafield and Alon Lavie. 2010. Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme. *The Prague Bulletin of Mathematical Linguistics*, 93:27–36.

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

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. In *Proc. of EACL*, pages 33–40.

Franz J. Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. In *Proc. of ACL*, pages 160–167.

Stephan Peitz et al. 2013. Joint WMT 2013 Submission of the QUAERO Project. In *Proc. of WMT*, pages 185–192.

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. of ICASSP*, volume 4, 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 *Proc. of AMTA*, pages 223–231.