The UZH System Combination System for WMT 2011

Rico Sennrich
Institute of Computational Linguistics
University of Zurich
Binzmühlestr. 14
CH-8050 Zürich
sennrich@cl.uzh.ch

Abstract

This paper describes the UZH system that was used for the WMT 2011 system combination shared task submission. We participated in the system combination task for the translation directions DE–EN and EN–DE. The system uses Moses as a backbone, with the outputs of the 2–3 best individual systems being integrated through additional phrase tables. The system compares well to other system combination submissions, with no other submission being significantly better. A BLEU-based comparison to the individual systems, however, indicates that it achieves no significant gains over the best individual system.

1 Introduction

For our submission to the WMT 2011 shared task, we built a system with the multi-engine MT approach described in (Sennrich, 2011), which builds on the idea by (Chen et al., 2007). A Moses SMT system (Koehn et al., 2007) is used as a backbone, trained on the WMT 2011 training data. Translation hypotheses by other systems are integrated through a second phrase table. In this second phrase table, the phrase translation probabilities and lexical weights are computed based on the word and phrase frequencies in both the translation hypotheses and a parallel training corpus. On the evaluation data in (Sennrich, 2011), this system significantly outperformed MEMT (Heafield and Lavie, 2010), which was among the best-performing system combination tools at WMT 2010 (Callison-Burch et al., 2010).

In this paper, we apply the same approach to a different translation scenario, namely the WMT 2011 shared task. We fail to significantly outperform the best individual system in terms of BLEU score. In section 2, we describe our system combination approach. In section 3, we present the results, and discuss possible reasons why the system fails to show the same performance gains as in the translation task on which it was evaluated initially.

2 System Description

We participated in the system combination task DE–EN and EN–DE. Since the combination is achieved by integrating translation hypotheses into an existing Moses system, which we will call the primary system, we first describe the methods and data used for training this primary system. Then, we describe how the translation hypotheses are selected out of the individual system submissions and integrated into the Moses system.

2.1 Primary System

For the training of the primary systems, we mostly followed the baseline instructions for the translation task[1]. We use *news-commentary* and *Europarl* as parallel training data. The language models are a linear interpolation of the *news-commentary*, *Europarl* and *news* corpora, optimized for minimal cross-entropy on the *newstest2008* data sets in the respective target language.

Additionally, we prune the primary phrase table using statistical significance tests, as described by (Johnson et al., 2007). For the translation direction DE–EN, the German source text is reordered based

[1] described at http://www.statmt.org/wmt11/baseline.html

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 166–170,
Edinburgh, Scotland, UK, July 30–31, 2011.

on syntactic parsing with Pro3GresDE (Sennrich et al., 2009), and reordering rules similar to those described by (Collins et al., 2005).

The Moses phrase table consists of five features: phrase translation probabilities in both translation directions ($p(\bar{t}|\bar{s})$ and $p(\bar{s}|\bar{t})$), lexical weights ($lex(\bar{t}|\bar{s})$ and $lex(\bar{s}|\bar{t})$), and a constant phrase penalty (Koehn et al., 2003). The computation of the phrase translation probabilities and lexical weights is based on the word, phrase and word/phrase pair frequencies that are extracted from the parallel corpus. We modified the Moses training scripts to collect and store these frequencies for later re-use.

We did not submit the primary system outputs to the Machine Translation shared task, since we did not experiment with new techniques. Instead, the primary system forms the backbone of the system combination system.

2.2 Integrating Secondary Phrase Tables

To combine the output of several systems, we train a second phrase table on the translation hypotheses of these systems. For this, we create a parallel corpus consisting of n translation hypotheses and n copies of the corresponding source text, both lowercased and detokenized.[2]

We compute the word alignment with MGIZA++ (Gao and Vogel, 2008), based on the word alignment model from the primary corpus that we have previously saved to disk.

After training a phrase table from the word-aligned corpus with Moses, the lexical weights and translation probabilities are rescored, using the sufficient statistics (i.e. the word, phrase and word/phrase pair counts) of both the primary and the secondary corpus. This rescoring step has been shown to markedly improve performance in (Sennrich, 2011). We will discuss its effects in section 3.1. The re-scored phrase table is integrated into the primary Moses system as an alternative decoding path, and tuned for maximal BLEU score on *newssyscombtune2011* with MERT.

[2]For convenience and speed, we combined the translation hypotheses for *newssyscombtune2011* and *newssyscombtest2011* into a single corpus. In principle, we could train separate phrase tables for each data set, or even for arbitrarily low numbers of sentences, without significant loss in performance (see (Sennrich, 2011)).

System	BLEU
Primary	21.11
Best individual	24.16
Submission	24.44
Vanilla scoring	24.42

Table 1: DE–EN results. Case-insensitive BLEU scores.

2.3 Hypothesis Selection

For the secondary phrase table, we chose to select the n best individual systems according to their BLEU score on the tuning set. We determined the optimal n empirically by trying different n, measuring each system's BLEU score on the tuning set and selecting the highest-scoring one. For the DE–EN translation task, $n = 2$ turned out to be optimal, for EN–DE, $n = 3$.

Chen et al. (2009) propose additional, tunable features in the phrase table to indicate the origin of phrase translations. For better comparability with the results described in (Sennrich, 2011), we did not add such features. This means that there are no *a priori* weights that bias the phrase selection for or against certain systems, but that decoding is purely driven by the usual Moses features: two phrase tables – the primary one and the re-scored, secondary one – the language model, the primary reordering model, and the corresponding parameters established through MERT.

3 Results

In the manual evaluation, the system combination submissions are only compared to each other, not to the individual systems. According to the manual evaluation, no other system combination submission outperforms ours by a statistically significant margin. In a comparison to individual systems, however, BLEU scores indicate that our system fails to yield a significant performance gain over the best individual system in this translation scenario.

In tables 1 and 2, we present case-insensitive BLEU scores (Papineni et al., 2002). As statistical significance test, we applied bootstrap resampling (Riezler and Maxwell, 2005). Tables 1 and 2 show the BLEU scores for the translation directions DE–EN and EN–DE, respectively. Systems included are the primary translation system described

System	BLEU
Primary	14.99
Best individual	17.44
Submission	17.51
Vanilla scoring	17.32

Table 2: EN–DE results. Case insensitive BLEU scores.

in section 2.1, the best individual system (online-B in both cases) and the submitted combination system. In terms of BLEU score, we achieved no statistically significant improvement over the best individual system.

As contrastive systems, we trained systems without the rescoring step described in section 2.2; we found no statistically significant difference from the submission system. In this translation task, the statistics from the parallel corpus seem to be ineffective at improving decoding, contrary to our findings in (Sennrich, 2011), where rescoring the phrase table improved BLEU scores by 0.7 points. We will address possible reasons for this discrepancy in the following section.

3.1 Interpretation

The main characteristic that sets our approach apart from other system combination software such as MANY (Barrault, 2010) and MEMT (Heafield and Lavie, 2010) is its reliance on word and phrase frequencies in a parallel corpus to guide decoding, whereas MANY and MEMT operate purely on the target side, without requiring/exploiting the source text or parallel data. We integrate the information from a parallel corpus into the decoding process by extracting phrase translations from the translation hypotheses and scoring these phrase translations on the basis of the frequencies from the parallel corpus.

The properties of this re-scored phrase table proved attractive for the translation task in (Sennrich, 2011), but less so for the WMT 2011 translation task. To explain why, let us look at $p(\overline{t}|\overline{s})$, i.e. the probability of a target phrase given a source phrase, as an example. It is computed as follows, c_{prim} and c_{sec} being the phrase count in the primary and secondary corpus, respectively.

$$p(\overline{t}|\overline{s}) = \frac{c_{prim}(\overline{s},\overline{t}) + c_{sec}(\overline{s},\overline{t})}{c_{prim}(\overline{s}) + c_{sec}(\overline{s})} \quad (1)$$

We can assume that $c_{sec}(\overline{s})$ and $c_{sec}(\overline{s},\overline{t})$ are mostly fixed, having values between 1 and the number of translation hypotheses.[3] If $c_{prim}(\overline{s})$ is high, the phrase translation probabilities in the secondary phrase table will only be marginally different from those in the primary phrase table (e.g. $\frac{500}{1000} = 0.5$ vs. $\frac{500+2}{1000+2} = 0.501$), whereas the secondary corpus has a stronger effect for phrases that are rare or unseen in the primary corpus (e.g. $\frac{1}{3} = 0.333$ vs. $\frac{1+2}{3+2} = 0.6$). Analogously, the same reasoning applies to $p(\overline{s}|\overline{t})$, $lex(\overline{t}|\overline{s})$ and $lex(\overline{s}|\overline{t})$.[4][5]

In short: the more frequent the phrases and phrase pairs in the primary corpus, the less effect does the secondary corpus have on the final feature values. This is a desirable behaviour if we can "trust" the phrase pairs extracted from the primary corpus. In (Sennrich, 2011), the primary corpus consisted of in-domain texts, whereas the translation hypotheses came from an out-of-domain SMT system and a rule-based one. There, it proved an effective strategy to only consider those translation hypotheses that either agreed with the data from the primary corpus, or for which the primary corpus had insufficient data, i.e. unknown or rare source words. With a primary system achieving a BLEU score of 17.18 and two translation hypotheses, scoring 13.29 and 12.94, we obtained a BLEU score of 20.06 for the combined system.

In the WMT 2011 system combination task, the statistics from the primary corpus failed to effectively improve translation quality. We offer these explanations based on an analysis of the results.

First, the 2–3 systems whose translation hypotheses we combine obtain higher scores than the primary system. This casts doubt on whether we should trust the scores from the primary system more than the translation hypotheses. And in fact, the results in table 1 and 2 show that the submission system

[3] Strictly speaking, this is only true if we build separate phrase tables for each sentence that is translated, and if there are no repeated phrases. This slight simplification serves illustrative purposes.

[4] For long phrases, phrase counts are typically low. Still, the primary corpus plays an important role in the computation of the lexical weights, which are computed from word frequencies, and thus typically less sparse than phrase frequencies.

[5] Rare target words may obtain a undesirably high probability, but are penalized in the language model. We set the LM log-probability of unknown words to -100.

(whose phrase table features take into account the primary corpus) is not better than a contrastive combination system with vanilla scoring, i.e. one that is solely based on the secondary corpus. We can also show why the primary corpus does not improve decoding by way of example. The German phrase *Bei der Wahl [der Matratze]* (English: *In the choice [of a mattress]*), is translated by the three systems as *in the selection*, *when choosing* and *in the election*. In this context, the last translation hypothesis is the least correct, but since the political domain is strongly represented in the training data, it is the most frequent one in the primary corpus, and the one being chosen by both the primary and the combined system.

Second, there seems to be a significant overlap in training data between the systems that we combine and our primary system[6]. We only saw few cases in which a system produced a translation against which there was evidence in our primary corpus. One instance is the German word *Kindergarten* (English: *kindergarten; nursery*), which is translated as *children's garden* by one system. In the combined system, this translation is dispreferred. (Chen et al., 2009) argue that a combination of dissimilar systems might yield better results. Rule-based systems could fulfill this role; they are also an attractive choice given their high quality (as judged by human evaluators) in earlier evaluations (e.g. WMT 2009 (Callison-Burch et al., 2009)). We did not pursue this idea, since we optimized for highest BLEU score, both during MERT and for the selection of the submission system, a scoring method that has been shown to undervalue rule-based systems (Callison-Burch et al., 2006).

The failure to outperform the individual best system in this translation task does not invalidate our approach. It merely highlights that different conditions call for different tools. Our approach relies strongly on parallel training data, in contrast to system combination tools such as MANY (Barrault, 2010) and MEMT (Heafield and Lavie, 2010). In this setting, this brought no benefit. However, when developing a SMT system for a specific domain and when combining an in-domain primary system with out-of-domain translation hypotheses, we expect that this strong dependence on the primary SMT system becomes an advantage. It allows the system to discriminate between source phrases that are well-documented in the primary training data, which will give other systems' hypotheses little effect, and those that occur rarely or not at all in the primary data, for which other systems may still produce a useful translation.

[6]This is especially true for all shared task participants building constrained systems. The amount of overlap between the anonymous online systems is unknown.

4 Conclusion

We described the UZH system combination submission to the Workshop of Machine Translation 2011. It uses the Moses architecture and includes translation hypotheses through a second phrase table. Its central characteristic is the extraction of phrase pairs from translations hypotheses and the scoring thereof on the basis of another parallel corpus. We find that, in the WMT 2011 system combination shared task, this approach fails to result in a significant improvement over the best individual system in terms of BLEU score. However, we argue that it is well suited for other translation tasks, such as the one described in (Sennrich, 2011).

Acknowledgments

This research was funded by the Swiss National Science Foundation under grant 105215_126999.

References

Loïc Barrault. 2010. MANY: Open source MT system combination at WMT'10. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 277–281, Uppsala, Sweden, July. Association for Computational Linguistics.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

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. Association for Computational Linguistics. Revised August 2010.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source decoder for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 193–196, Morristown, NJ, USA. Association for Computational Linguistics.

Yu Chen, Michael Jellinghaus, Andreas Eisele, Yi Zhang, Sabine Hunsicker, Silke Theison, Christian Federmann, and Hans Uszkoreit. 2009. Combining multi-engine translations with Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 42–46, Morristown, NJ, USA. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 531–540, Morristown, NJ, USA. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2010. CMU multi-engine machine translation for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 301–306, Stroudsburg, PA, USA. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007*, pages 177–180, Prague, Czech Republic, June.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A New Hybrid Dependency Parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*, Potsdam, Germany.

Rico Sennrich. 2011. Combining multi-engine machine translation and online learning through dynamic phrase tables. In *15th Annual Conference of the European Association for Machine Translation (EAMT 2011)*, Leuven, Belgium.