DFKI's SMT System for WMT 2012

David Vilar
German Research Center for Artificial Intelligence (DFKI GmbH)
Language Technology Lab
Berlin, Germany
david.vilar@dfki.de

Abstract

We describe DFKI's statistical based submission to the 2012 WMT evaluation. The submission is based on the freely available machine translation toolkit Jane, which supports phrase-based and hierarchical phrase-based translation models. Different setups have been tested and combined using a sentence selection method.

1 Introduction

In this paper we present DFKI's submission for the 2012 MT shared task based on statistical approaches. We use a variety of phrase-based and hierarchical phrase-based translation systems with different configurations and enhancements and compare their performance. The output of the systems are later combined using a sentence selection mechanism. Somewhat disappointingly the sentence selection hardly improves over the best single system.

DFKI participated in the German to English and English to German translation tasks. Technical problems however hindered a more complete system for this last translation direction.

This paper is organized as follows: Section 2 reports on the different single systems that we built for this shared task. Section 3 describes the sentence selection mechanism used for combining the output of the different systems. Section 4 concludes the paper.

2 Single Systems

For all our setups we used the Jane toolkit (Vilar et al., 2010a), which in its current version supports both phrase-based and hierarchical phrase-based translation models. In this Section we present the different settings that we used for the task.

The bilingual training data used for training all systems was the combination of the provided Europarl and News data. We also used two baseline 4-gram language models trained on the same Europarl training data and on the enhanced News Commentary monolingual training data. The newstest2010 dataset was used for optimization of the systems.

2.1 Phrase-based System

The first system is a baseline phrase-based system trained on the available bilingual training data. Word alignments is trained using GIZA++ (Och and Ney, 2003), phrase extraction is performed with Jane using standard settings, i.e. maximum source phrase length 6, maximum target phrase length 12, count features, etc. Consult the Jane documentation for more details. For reordering the standard distance-based reordering model is computed. Scaling factors are trained using MERT on n-best lists.

2.1.1 Verb reorderings

Following (Popović and Ney, 2006), for German to English translation, we perform verb reordering by first POS-tagging the source sentence and afterwards applying hand-defined rules. This includes rules for reordering verbs in subordinate clauses and participles.

2.1.2 Moore LM

Moore and Lewis (2010) propose a method for filtering large quantities of out-of-domain language-model training data by comparing the cross-entropy

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 382–387,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

of an in-domain language model and an out-of-domain language model trained on a random sampling of the data. We followed this approach to filter the news-crawl corpora provided the organizers. By experimenting on the development set we decided to use a 4-gram language model trained on 15M filtered sentences (the original data comprising over 30M sentences).

2.2 Hierarchical System

We also trained a hierarchical system on the same data as the phrase-based system, and also tried the additional language model trained according to Section 2.1.2, as well as the verb reorderings described in Section 2.1.1.

2.2.1 Poor Man's Syntax

Vilar et al. (2010b) propose a "syntax-based" approach similar to (Venugopal et al., 2009), but using automatic clustering methods instead of linguistic parsing for defining the non-terminals used in the resulting grammar. The main idea of the method is to cluster the words (mimicking the concept of Part-of-Speech tagging), performing a phrase extraction pass using the word classes instead of the actual words and performing another clustering on the phrase level (corresponding to the linguistic classes in a parse tree).

2.2.2 Lightly-Supervised Training

Huck et al. (2011) propose to augment the monolingual training data by translating available additional monolingual data with an existing translation system. We adapt this approach by translating the data selected according to Section 2.1.2 with the phrase-based translation system described in Section 2.1, and use this additional data to expand the bilingual data available for training the hierarchical phrase-based system.[1]

2.3 Experimental Results

Table 1 shows the results obtained for the German to English translation direction on the newstest2011 dataset. The baseline phrase-based system obtains a BLEU score of 18.2%. The verb reorderings achieve an improvement of 0.6% BLEU, and adding the additional language model obtains an additional 1.6% BLEU improvement.

The hierarchical system baseline achieves a better BLEU score than the baseline PBT system, and is comparable to the PBT system with additional reorderings. In fact, adding the verb reorderings to the hierarchical system slightly degrades its performance. This indicates that the hierarchical model is able to reflect the verb reorderings necessary for this translation direction. Adding the bigger language model of Section 2.1.2 also obtains a nice improvement of 1.4% BLEU for this system. On the other hand and somewhat disappointingly, the lightly supervised training and the poor man's syntax approach are not able to improve translation quality.

For the English to German translation direction we encountered some technical problems, and we were not able to perform as many experiments as for the opposite direction. The results are shown in Table 2 and show similar trends as for the German to English direction, except that the hierarchical system in this case does not outperform the PBT baseline.

3 Sentence Selection

In this section we will describe the system combination method based on sentence selection that we used for combining the output of the systems described in Section 2. This approach was tried successfully in (Vilar et al., 2011).

We use a log-linear model for computing the scores of the different translation hypotheses, generated by all the systems described in Section 2, i.e. those listed in Tables 1 and 2. The model scaling factors are computed using a standard MERT run on the newstest2011 dataset, optimizing for BLEU. This is comparable to the usual approach used for rescoring n-best lists generated by a single system, and has been used previously for sentence selection purposes (see (Hildebrand and Vogel, 2008) which uses a very similar approach to our own). Note that no system dependent features like translation probabilities were computed, as we wanted to keep the system general.

We will list the features we compute for each of

[1] The decision of which system to use to produce the additional training material follows mainly a practical reason. As the hierarchical model is more costly to train and at decoding time, we chose the phrase-based system as the generating system.

System	BLEU[%]
PBT Baseline	18.2
PBT + Reordering	18.8
PBT + Reordering + Moore LM	20.4
Hierarchical Baseline	18.7
Hierarchical + Moore LM	20.1
Hierarchical + Moore LM + Lightly Supervised	19.8
Poor Man's Syntax	18.6
Hierarchical + Reordering	18.5

Table 1: Translation results for the different single systems, German to English.

System	BLEU[%]
PBT Baseline	12.4
Hierarchical Baseline	11.6
Hierarchical + Moore LM	13.1
Poor Man's Syntax	11.6

Table 2: Translation results for the different single systems, English to German

the systems. We have used features that try to focus on characteristics that humans may use to evaluate a system.

3.1 Cross System BLEU

BLEU was introduced in (Papineni et al., 2002) and it has been shown to have a high correlation with human judgement. In spite of its shortcomings (Callison-Burch et al., 2006), it has been considered the standard automatic measure in the development of SMT systems (with new measures being added to it, but not substituting it, see for e.g. (Cer et al., 2010)).

Of course, the main problem of using the BLEU score as a feature for sentence selection in a real-life scenario is that we do not have the references available. We overcame this issue by generating a custom set of references for each system, using the other systems as gold translations. This is of course inexact, but n-grams that appear on the output of different systems can be expected to be more probable to be correct, and BLEU calculated this way gives us a measure of this agreement. This approach can be considered related to n-gram posteriors (Zens and Ney, 2006) or minimum Bayes risk decoding (e.g. (Ehling et al., 2007)) in the context of n-best rescoring, but applied without prior weighting (unavailable directly) and more focused on the evaluation interpretation.

We generated two features based on this idea. The first one is computed at the system level, i.e. it is the same for each sentence produced by a system and serves as a kind of prior weight similar to the one used in other system combination methods (e.g. (Matusov et al., 2008)). The other feature was computed at the sentence level. For this we used the smoothed version of BLEU proposed in (Lin and Och, 2004), again using the output of the rest of the systems as pseudo-reference. As optimization on BLEU often tends to generate short translations, we also include a word penalty feature.

3.2 Error Analysis Features

It is safe to assume that a human judge will try to choose those translations which contain the least amount of errors, both in terms of content and grammaticality. A classification of errors for machine translation systems has been proposed in (Vilar et al., 2006), and (Popović and Ney, 2011) presents how to compute a subset of these error categories automatically. The basic idea is to extend the familiar Word Error Rate (WER) and Position independent word Error Rate (PER) measures on word and base-form[2] levels to identify the different kind of errors. For our system we included following features:

Extra Word Errors (EXTer) Extra words in the hypothesis not present in the references.

Inflection Errors (hINFer) Words with wrong inflection. Computed comparing word-level errors and base-form-level errors.

Lexical Errors (hLEXer) Wrong lexical choices in the hypothesis with respect to the references.

Reordering Errors (hRer) Wrong word order in the hypothesis.

Missing Words (MISer) Words present in the reference that are missing in the hypothesis.

All these features are computed using the open source Hjerson[3] tool (Popović, 2011), which also outputs the standard WER metric, which we added as an additional feature.

As was the case in Section 3.1, for computing these measures we do not have a reference available, and thus we use the rest of the systems as pseudo-references. This has the interesting effect that some "errors" are actually beneficial for the performance of the system. For example, it is known that systems optimised on the BLEU metric tend to produce short hypotheses. In this sense, the extra words considered as errors by the EXTer measure may be actually beneficial for the overall performance of the system.

3.3 IBM1 Scores

IBM1-like scores on the sentence level are known to perform well for the rescoring of n-best lists from a single system (see e.g. (Hasan et al., 2007)). Additionally, they have been shown in (Popovic et al., 2011) to correlate well with human judgement for evaluation purposes. We thus include them as additional features.

	De-En	En-De
Best System	20.4	13.1
Worst System	18.2	11.6
Sentence Selection	20.9	13.3

Table 3: Sentence selection results

3.4 Additional Language Model

We used a 5-gram language model trained on the whole news-crawl corpus as an additional model for rescoring. We used a different language model as the one described in Section 2.1.2 as not to favor those systems that already included it at decoding time.

3.5 Experimental Results

The sentence selection improved a little bit over the best single system for German to English translation, but hardly so for English to German, as shown in Table 3. For English to German this can be due to the small amount of systems that were available for the sentence selection system. Note also that these results are measured on the same corpus the system was trained on, so we expect the improvement on unseen test data to be even smaller. Nevertheless the sentence selection system constituted our final submission for the MT task.

4 Conclusions

For this year's evaluation DFKI used a statistical system based around the Jane machine translation toolkit (Vilar et al., 2010a), working in its two modalities: phrase-based and hierarchical phrase-based models. Different enhancements were tried in addition to the baseline configuration: POS-based verb reordering, monolingual data selection, poor man's syntax and lightly supervised training, with mixed results.

A sentence selection mechanism has later been applied in order to combine the output of the different configurations. Although encouraging results had been obtained in (Vilar et al., 2011), for this task we found only a small improvement. This may be due to the strong similarity of the systems, as they are basically trained on the same data. In (Vilar et al., 2011) the training data was varied across the systems, which may have produced a bigger variety in

[2] Computed using the TreeTagger tool (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)

[3] The abbreviations for the features are taken over directly from the output of the tool.

the translation outputs that can be of advantage for the selection mechanism. This is an issue that should be explored in more detail for further work.

We also plan to do a comparison with system combination approaches where new hypotheses can be generated (instead of selecting one from a predefined set), and study under which conditions each approach is more suited than the other.

5 Acknowledgements

This work was done with the support of the TaraXŰ Project[4], financed by TSB Technologiestiftung Berlin-Zukunftsfonds Berlin, co-financed by the European Union-European fund for regional development.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 555–563, Los Angeles, CA, USA.

Nicola Ehling, Richard Zens, and Hermann Ney. 2007. Minimum Bayes risk decoding for BLEU. In *Annual Meeting of the Assoc. for Computational Linguistics*, pages 101–104, Prague, Czech Republic, June.

Saša Hasan, Richard Zens, and Hermann Ney. 2007. Are very large N-best lists useful for SMT? In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*, pages 57–60, Rochester, NY, April. Association for Computational Linguistics.

A.S. Hildebrand and S. Vogel. 2008. Combination of machine translation systems via hypothesis selection from combined n-best lists. In *MT at work: Proc. of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 254–261.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *The EMNLP 2011 Workshop on Unsupervised Learning in NLP*, Edinburgh, UK, July.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proc. of the 20th international conference on Computational Linguistics*, COLING '04, Geneva, Switzerland.

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System combination for machine translation of spoken and written language. *IEEE Transactions on Audio, Speech and Language Processing*, 16(7):1222–1237, September.

R.C. Moore and W. Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

Maja Popović and Hermann Ney. 2006. POS-based word reorderings for statistical machine translation. In *International Conference on Language Resources and Evaluation*, pages 1278–1283, Genoa, Italy, May.

Maja Popović and Hermann Ney. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4):657–688, December.

Maja Popovic, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: Ibm1 scores as evaluation metrics. In *Proc. of the Sixth Workshop on Statistical Machine Translation*, pages 99–103. Association for Computational Linguistics, July.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, pages 59–68.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference Grammars: Softening Syntactic Constraints to Improve Statistical Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–244, Boulder, Colorado, USA, June.

[4] http://taraxu.dfki.de

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010a. Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon Models. In *Proc. of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 262–270, Uppsala, Sweden, July.

David Vilar, Daniel Stein, Stephan Peitz, and Hermann Ney. 2010b. If I Only Had a Parser: Poor Man's Syntax for Hierarchical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 345–352, Paris, France, December.

David Vilar, Eleftherios Avramidis, Maja Popović, and Sabine Hunsicker. 2011. Dfki's sc and mt submissions to iwslt 2011. In *International Workshop on Spoken Language Translation*, San Francisco, CA, USA, December.

R. Zens and H. Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Workshop on Statistical Machine Translation*, pages 72–77, New York City, June.