DFKI Hybrid Machine Translation System for WMT 2011 - On the Integration of SMT and RBMT

Jia Xu and Hans Uszkoreit and Casey Kennington and David Vilar and Xiaojun Zhang
DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken Germany
{Jia.Xu,uszkoreit,David.Vilar}@dfki.de, {bakuzen,xiaojun.zhang.iiken}@gmail.com

Abstract

We present the DFKI hybrid translation system at the WMT workshop 2011. Three SMT and two RBMT systems are combined at the level of the final translation output. The translation results show that our hybrid system significantly outperformed individual systems by exploring strengths of both rule-based and statistical translations.

1 Introduction

Machine translation (MT), in particular the statistical approach to it, has undergone incremental improvements in recent years. While rule-based machine translation (RBMT) maintains competitiveness in human evaluations. Combining the advantages of both approaches have been investigated by many researchers such as (Eisele et al., 2008). Nonetheless, significant improvements over statistical approaches still remain to be shown. In this paper, we present the DFKI hybrid system in the WMT workshop 2011. Our system is different from the system of the last year (Federmann et al., 2010), which is based on the shallow phrase substitution. In this work, two rule-based translation systems are applied. In addition, three statistical machine translation systems are built, including a phrase-based, a hierarchical phrase-based and a syntax-based system. Instead of combining with rules or post-editing, we perform system combination on the final translation hypotheses. We applied the CMU open toolkit (Heafield and Lavie, 2010) among numerous combination methods such as (Matusov, 2009), (Sim et al., 2007) and (He et al., 2008). The final translation output outperforms each individual output significantly.

2 Individual translation systems

2.1 Phrase-based system

We use the IBM model 1 and 4 (Brown et al., 1993) and Hidden-Markov model (HMM) (Vogel et al., 1996) to train the word alignment using the mgiza toolkit[1]. We applied the EMS in Moses (Koehn et al., 2007) to build up the phrase-based translation system. Features in the log-linear model include translation models in two directions, a language model, a distortion model and a sentence length penalty. A dynamic programming beam search algorithm is used to generate the translation hypothesis with maximum probability. We applied a 5-gram mixture language model with each sub-model trained on one fifth of the monolingual corpus with Kneser-Ney smoothing using SRILM toolkit (Stolcke, 2002). We did not perform any tuning, because it hurts the evaluation performance in our experiments.

2.2 Syntax-based system

To capture the syntactic structure, we also built a tree-based system using the same configuration of EMS in Moses (Koehn et al., 2007). Tree-based models operate on so-called grammar rules, which include variables in the mapping rules. To increase the diversity of models in combination, the language model in each individual translation system is trained differently. For the tree-based system, we applied a 4-gram language model with Kneser-Ney smoothing using SRILM toolkit (Stolcke, 2002) trained on the whole monolingual corpus. The test2007 news part is applied to tune the feature weights using mert, because the tuning on test2007

[1] http://geek.kyloo.net/software/doku.php/mgiza:overview

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 485–489,
Edinburgh, Scotland, UK, July 30–31, 2011.

improves the translation performance more than the tuning on test2008 in a small-scale experiment for the tree-based system.

2.3 Hierarchical phrase-based system

For the hierarchical system, we used the open source hierarchical phrased-based system Jane, developed at RWTH and free for non-commercial use (Vilar et al., 2010). This approach is an extension of the phrase-based approach, where the phrases are allowed to have gaps (Chiang, 2007). In this way long-range dependencies and reorderings can be modeled in a consistent statistical framework.

The system uses a fairly standard setup, trained using the bilingual data provided by the organizers, word aligned using the mgiza. Two 5-gram language models were used during decoding: one trained on the monolingual part of the bilingual training data, and a larger one trained on the additional news data. Decoding was carried out using the cube pruning algorithm. The tuning is performed on test2008 without further experiments.

2.4 Rule-based systems

We applied two rule-based translation systems, the Lucy system (Lucy, 2011) and the Linguatec system (Aleksić and Thurmair, 2011). The Lucy system is a recent offspring of METAL. The Linguatec system is a modular system consisting of grammar, lexicon and morphological analyzers based on logic programming using slot grammar.

3 Hybrid translation

A hybrid approach combining rule-based and statistical machine translation is usually investigated with an in-box integration, such as multi-way translation (Eisele et al., 2008), post-editing (Ueffing et al., 2008) or noun phrase substitution (Federmann et al., 2010). However, significant improvements over state-of-the-art statistical machine translation are still expected. In the meanwhile system combination methods for instance described in (Matusov, 2009), (Sim et al., 2007) and (He et al., 2008) are mostly evaluated to combine statistical translation systems, rule-based systems are not considered. In this work, we integrate the rule-based and statistical machine translation system on the level of the final translation hypothesis and treat the rule-based system anonymously as an individual system. In this way an black-box integration is allowed using the current system combination techniques.

	PBT	Syntax
PBT-2010	18.32	
Max80words	20.65	21.10
Max100words	20.78	
+Compound	21.52	22.13
+Newparallel	21.77	

Table 1: Translation performance BLEU[%] on phrase/syntax-based system using various settings evaluated on test10.

We applied the CMU open toolkit (Heafield and Lavie, 2010) MEMT, a package by Kenneth Heafield to combine the translation hypotheses. The language model is trained on the target side of the parallel training corpus using SRILM (Stolcke, 2002). We used only the Europarl part to train language models for tuning and all target side of parallel data to train language models for decoding. The beam size is set to 80, and 300 nbest is considered.

4 Translation experiments

4.1 MT Setup

The parallel training corpus consists of 1.8 million German-English parallel sentences from Europarl-v6 (Koehn, MT Summit 2005) and news-commentary with 48 million tokenized German words and 54 million tokenized English words respectively. The monolingual training corpus contains the target side of the parallel training corpus and the additional monolingual language model training data downloaded from (SMT, 2011). We did not apply the large-scale Gigaword corpus, because it does not significantly reduce the perplexity of our language model but raises the computational requirement heavily.

4.2 Single systems

For each individual translation system, different configurations are experimented to achieve a higher translation quality. We take phrase- and syntax-based translation system as examples. Table 1 presents official submission result on DE-EN by

PBT+Syntax	20.37
PBT+Syntax+HPBT	20.78
PBT+HPBT+Linguatec+Lucy	20.27
PBT+Syntax+HPBT+Linguatec+Lucy	20.81

Table 2: Translation performance BLEU[%] on test2011 using hybrid system tuned on test10 with various settings (DE-EN).

	Test10	Test08	Test11
Hybrid-2010	17.43		
PBT	21.77	20.70	20.40
Syntax	22.13	20.50	20.49
HPBT	19.21	18.26	17.06
Linguatec	16.59	16.07	15.97
Lucy	16.57	16.66	16.68
Hybrid-2011	23.88	21.13	21.25

Table 3: Translation performance BLEU[%] on three test sets using different translation systems in 2011 submission (DE-EN).

	Test10	Test11
Hybrid-2010	14.42	
PBT	15.46	14.05
Linguatec	14.92	12.92
Lucy	13.77	13.0
Hybrid-2011	15.55	15.83

Table 4: Translation performance BLEU[%] on two test sets using different translation systems in 2011 submission (EN-DE).

DFKI in 2010. In 2010's translation system only Europarl parallel corpus was applied, and the translation output was evaluated as 18.32% in the BLEU score. In 2011, we added the News Commentary parallel corpus and trained the language model on all monolingual data provided by (SMT, 2011) except for Gigaword. As shown in Table 1, if we increase the maximum sentence length of the training corpus from 80 to 100, the BLEU score increases from 20.65% to 20.78%. In the error analysis, we found that many OOVs come from the compound words in German. Therefore, we applied the compound splitting for both German and English by activating the corrensponding settings in the EMS in Moses. This leads to a further improvement of nearly 1% in the BLEU score. As we add the new parallel corpus provided on the homepage of SMT workshop in 2011 (SMT, 2011) to the corpus in 2010, a slight improvement can be achieved. Within one year, the score for the DFKI PBT system DE-EN has improved by nearly 3.5% absolute and 20% relative BLEU score points, as shown in Table 1.

In the phrase-based translation, the tuning was not applied, because it improves the results on the held-out data but hurts the results on the evaluation set. In our observation, the decrease is in the range of 0.01% to 1% in the BLEU score. However tuning does help for the Tree-based system. Therefore we applied the test2007 to optimize the parameters, which enhanced the BLEU score from 17.52% to 21.10%. The compound splitting also improves the syntax system, with about 1% in the BLEU score. We did not add the new parallel corpus into the training for syntax system due to its larger computational requirement than that of the phrase-based system.

4.3 Hybrid system

We applied test10 as the held-out data to tune the German-English and English-German translation systems. For experiments, we applied a small-scaled 4-gram language model trained only on the target side of the Europarl parallel training data. As shown in Table 2, different combinations are performed on the hypotheses generated from single systems. We first combined the PBT with syntax system, then together with the HPBT system. The translation result in the BLEU score performs best when we combine all three statistical machine translation systems and two rule-based systems together.

4.4 Evaluation results

For the decoding during the WMT evaluation, we applied a larger 4-gram language model trained on the target side of all parallel training corpus. As shown in Table 3, in last year's evaluation the DFKI hybrid translation result was evaluated as 17.34% in the BLEU score. In 2011, among all the translation systems, the syntax system performs the best on test10 and test11, while the PBT performs the

SRC	Diese Verordnung wurde vom Gesundheitsministerium in diesem Jahr einigermassen gemildert - die Kühlschrankpflicht fiel weg.
REF	It was mitigated by the Ministry of Health this year - the obligation to have a refrigerator has been removed.
PBT	This regulation by the Ministry of Health in this year - somewhat mitigated the fridge duty fell away.
Syntax	This regulation was somewhat mitigated by the Ministry of Health this year - the refrigerator duty fell away.
HPBT	This regulation was by the Ministry of Health in reasonably Dokvadze this year - the Kühlschrankpflicht fell away.
Linguatec	This ordinance was soothed to some extent by the brazilian ministry of health this year, the refrigerator duty was discontinued.
Lucy	This regulation was quite moderated by the Department of Health, Education and Welfare this year - the refrigerator duty was omitted.
Hybrid	This regulation was somewhat mitigated by the Ministry of Health this year - the fridge duty fell away.
SRC	Die Deregulierung und Bakalas ehemalige Bergarbeiterwohnungen sind ein brisantes Thema.
REF	Deregulation and Bakala 's former mining flats are local hot topic.
PBT	The deregulation and Bakalas former miners' homes are a sensitive issue.
Syntax	The deregulation and Bakalas former miners' homes are a sensitive issue.
HPBT	The deregulation and Bakalas former Bergarbeiterwohnungen are a hot topic.
Linguatec	Former miner flats are an explosive topic the deregulation and Bakalas.
HPBT	The deregulation and Bakalas former miner apartments are an explosive topic.
Hybrid	The deregulation and Bakalas former miners' apartments are a sensitive issue.

Table 5: Examples of translation output by the different systems.

best on test08. The rule-based sytems, Linguatec and Lucy are expected to have a higher score in the human evaluation than in the automatic evaluation. Furthermore, as we can see from Table 3, there is still room to improve the Jane system, with better modeling, configurations or even higher-order language model. Using the hybrid system we successfully improved the translation result to 23.88% on test10. The hybrid system outperforms the best single system by 0.43% and 0.76% in the BLEU score on the test08 and test11, respectively.

For the translation from English to German, the translation result of last year's submission was evaluated as 14.42% in the BLEU score, as shown in Table 4. In this year, the phrase-based translation result is 15.46% in the BLEU score. We only set up one statistical translation system due to time limitation. With the respect of the BLEU score, phrase-based translation outperforms rule-based translations. Between rule-based translation systems, Linguatec performs better on the test10 (14.92%) and Lucy performs better on the test11 (13.0%). Combining three translation hypotheses leads to a smaller improvement (from 15.46% to 15.55%) on the test10 and a greater improvement (from 14.05% to 15.83%) on the test11 in the BLEU score over the single best translation system. Comparing to last year's translation output, the improvement is over one percent absolutely (from 14.42% to 15.55%) in the BLEU score on the test10.

4.5 Output examples

Table 5 shows two translation examples from the MT output of the test2011. We list the source sentence in German and its reference translation as well as the translation results generated by different translation systems. As can be seen from Table 5, the translation quality of source sentences is greatly improved using the hybrid system over the single individual systems. Translations of words and word orderings are more appropriate by the hybrid system.

5 Conclusion and future work

We presented the DFKI hybrid translation system submitted in the WMT workshop 2011. The hybrid translation is performed on the final translation output by individual systems, including a phrase-based system, a syntax-based system, a hierarchical phrase-based system and two rule-based systems. Combining the results from statistical and rule-based systems significantly improved the translation performance over the single-best system, which is shown by the automatic evaluation scores and the output examples. Despite of the encouraging results, there is still room to improve our system, such as the tuning in the phrase-based translation and a better language model in the combination.

References

Vera Aleksić and Gregor Thurmair. 2011. Personal translator at wmt2011 - a rule-based mt system with hybrid components. In *Proceedings of WMT workshop*.

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

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

Andreas Eisele, Christian Federmann, Hans Uszkoreit, Hervé Saint-Amand, Martin Kay, Michael Jellinghaus, Sabine Hunsicker, Teresa Herrmann, and Yu Chen. 2008. Hybrid architectures for multi-engine machine translation. In *Proceedings of Translating and the Computer 30*, pages ASLIB, ASLIB/IMI, London, United Kingdom, November.

Christian Federmann, Andreas Eisele, Hans Uszkoreit, Yu Chen, Sabine Hunsicker, and Jia Xu. 2010. Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 237–248, Uppsala, Sweden. John Benjamins.

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 EMNLP*, October.

Kenneth Heafield and Alon Lavie. 2010. Voting on n-grams for machine translation system combination. In *Proc. Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado, October.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

Philipp Koehn. MT Summit 2005. Europarl: A parallel corpus for statistical machine translation.

Lucy. 2011. Home page of software lucy and services. `http://www.lucysoftware.com`.

Evgeny Matusov. 2009. *Combining Natural Language Processing Systems to Improve Machine Translation of Speech*. Ph.D. thesis, Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD.

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 *IN IEEE INT. CONF. ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*.

SMT. 2011. Sixth workshop on statistical machine translation home page. http://www.statmt.org/wmt11/.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference On Spoken Language Processing*, pages 901–904, Denver, Colorado, September.

Nicola Ueffing, Jens Stephan, Evgeny Matusov, Lo ic Dugast, George F. Foster, Roland Kuhn, Jean Senellart, and Jin Yang. 2008. Tighter integration of rule-based and statistical mt in serial system combination. In *Proceedings of COLING 2008*, pages 913–920.

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

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.