DFKI's system for WMT16 IT-domain task, including analysis of systematic errors

Eleftherios Avramidis, Aljoscha Burchardt, Vivien Macketanz and Ankit Srivastava
German Research Center for Artificial intelligence (DFKI)
Language Technology Lab, Berlin
firstname.lastname@dfki.de

Abstract

We are presenting a hybrid MT approach in the WMT2016 Shared Translation Task for the IT-Domain. Our work consists of several translation components based on rule-based and statistical approaches that feed into an informed selection mechanism. Additions to last year's submission include a WSD component, a syntactically-enhanced component and several improvements to the rule-based component, relevant to the particular domain. We also present detailed human evaluation on the output of all translation components, focusing on particular systematic errors.

1 Introduction

We are presenting extensions on our hybrid MT approach from the WMT 2015 translation task in the generic-domain (Avramidis et al., 2015). The system combines several SMT and RBMT components that feed into an informed selection mechanism. For WMT 2016, several new system components have been submitted to the IT-task that are described in more detail in this paper.

In our work, detailed evaluation of translation quality using a wide variety of methods from automatic scores to human error annotation is an active part of the MT development process. Already in previous work (Popović et al., 2014), we have argued for an approach to MT research and development (R&D) that makes a more direct use of the knowledge and expertise of language professionals.

One of the reasons is that it is difficult to build hybrid architectures (that take advantage of the fact that different engines make different errors) solely based on the rough feedback provided by automatic scores. As scores like BLEU (Papineni et al., 2002) are not suitable for comparison across different types of engines like Statistical Machine Translation (SMT) and Rule-based Machine Translation (RBMT), we have included human feedback by a language professional in the development of the components reported in this paper.

To this end, we complement our system development with specific manual analysis. We have identified and manually inspected phenomena in the given domain that frequently lead to errors in our engines.

We are using the insights gained from this detailed analysis to guide further improvements of our engines and selection mechanism, some of which are detailed below. Therefore, the components developed follow the direction of addressing some of the most observed systematic issues. Nevertheless, the systems submitted to this task are only a stage in the continuous development effort.

The short paper is structured as follows: Section 2 includes a description of the individual components and the hybridization mechanism, section 3 presents a detailed manual evaluation focusing on systematic errors, whereas conclusions and ideas for further work are given in section 4.

2 System components

We hereby present the systems that appear in our submissions and our hybrid system:

2.1 Phrase-based SMT baseline

The baseline system consists of a basic phrase-based SMT model, trained with the state-of-the-art settings on both the generic and technical data. The translation table was trained on a concatenation of generic and technical data, filtering out the sentences longer than 80 words. Batch 1 was used as a tuning set for MERT (Och, 2003).

One language model (monolingual) of order 5 was trained on the target side from both the

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 415–422,
Berlin, Germany, August 11-12, 2016.

corpus	entries	words
Chromium browser	6.3K	55.1K
Drupal	4.7K	57.4K
Libreoffice help	46.8K	1.1M
Libreoffice UI	35.6K	143.7K
Ubuntu Saucy	182.9K	1.6M
Europarl (mono)	2.2M	54.0M
News (mono)	89M	1.7B
Commoncrawl (parallel)	2.4M	53.6M
Europarl (parallel)	1.9M	50.1M
MultiUN (parallel)	167.6K	5.8M
News Crawl (parallel)	201.3K	5.1M

Table 1: Size of corpora used for SMT.

technical (IT-domain) and Europarl corpora, plus one language model was trained on the target-language news corpus from the years 2007 to 2013 (Callison-Burch et al., 2007). All language models were interpolated on the tuning set (Schwenk and Koehn, 2008). The size of the training data is shown in Table 1.

The text has been tokenized and truecased (Koehn et al., 2008) prior to the training and the decoding, and de-tokenized and de-truecased afterwards. A few regular expressions were added to the tokenizer, so that URLs are not tokenized before being translated. Normalization of punctuation was also included, mainly in order to fix several issues with variable typography on quotes.

The phrase-based SMT system was trained with Moses (Koehn, 2010) using EMS (Koehn, 2010), whereas the language models were trained with SRILM (Stolcke, 2002) and queried with KenLM (Heafield, 2011).

All statistical systems presented below are extensions of this system, also based on the same data and settings, unless stated otherwise.

2.2 SMT with Word Sense Disambiguation

The word-sense-disambiguated SMT system is a factored phrase-based statistical system with two decoding paths, one basic and one alternative. In the basic path, all nouns of the source language (English) have been annotated with a WSD system (Weissenborn et al., 2015) that assigns BabelNet senses to nouns and has recently shown improvements over state-of-the-art results on several corpora. The sense labels are estimated based on the disambiguation analysis on the sentence level by choosing the best ranked sense out of the ones provided by the WSD system. Each produced WSD label replaces the respective base word form of the noun. In the alternative path, non-annotated input is used. The alternative path allows for decoding phrases when there are no WSD labels or the decoder cannot form a translation with a good probability.

	system variants	BLEU	METEOR
1.	SMT baseline	31.06	55.8
2.	sense → word	25.52	50.4
3.*	sense → word, word → word (alt)	29.89	54.8
4.	word → word, sense → word (alt)	29.88	54.3

Table 2: Automatic scores for factored SMT variants with WSD. (*) indicates the version included in the selection mechanism.

Due to the high computational demands of the WSD annotation, this model was trained on less data than the respective phrase-based models, using the first 1.1M sentences of Europarl and ommiting the entire Commoncrawl. We experimented with four different settings concerning the translation path. These settings with the corresponding automatic scores are depicted in Table 2, which includes the results on the development set 2. On this set, WSD does not show a positive effect over the baseline in terms of automatic scores.

2.3 Syntax-enhanced SMT

Motivated by the importance of grammar in the translation between English and German, we developed a syntax-enhanced SMT system. The process is similar to that of our baseline, but this version includes syntax-aware phrase extraction. Phrase pairs in the baseline SMT system were augmented with linguistically-motivated phrase pairs. These phrases were extracted by generating constituency and dependency parse trees for both the source and target languages, followed by node-aligning the parallel parse trees using a statistical tree aligner (Zhechev, 2009). The syntax-aware phrase extraction algorithm obtains surface-level chunks (syntax-aware) from the aligned subtrees (Srivastava and Way, 2009).

Intermediate experiments were conducted by using either constituency parsing or dependency

parsing and it was discovered that despite containing phrase pairs unique to each parsing model (around 28%), no statistically significant difference was observed in the MT system performance. We therefore present the version that uses both of them by concatenating all phrase pairs in one table in an attempt to beenfit from multiple knowledge sources (Srivastava et al., 2009). Aditionally informed by the manual inspection in Section 3, we performed a pseudo-Named Entity Recognition (words and phrases tagged as nouns) in order to identify in-domain terminology and translate them separately in a post-decoding automatic post-editing framework.

For the constituency and dependency parsing we employed the Berkeley Parser (Petrov and Klein, 2007) and the Stanford Dependency Parser (Klein and Manning, 2003) respectively.

2.4 Rule-based component

The rule-based system Lucy (Alonso and Thurmair, 2003) is also part of our experiment, due to its state-of-the-art performance in the previous years. Additionally, manual inspection on the development set has shown that it provides better handling of complex grammatical phenomena particularly when translating into German, due to the fact that it operates based on transfer rules from the source to the target syntax tree.

This year's work on RBMT focuses on issues revealed through manual inspection of its performance on the development set:

- **Separate menu items**: The rule-based system was observed to be incapable of handling menu items properly, mostly when they were separated by the ">" symbol, as they often ended up as compounds. We identified the menu items by searching for consequent title-cased chunks before and after each separator. These items were translated separately from the rest of the sentence, to avoid them being bundled as compounds. The rule-based system was then forced to treat the pre-translated menu items as chunks that should not be translated.
- **Menu items by SMT**: Additionally, we used the method above to check whether menu items could be translated with the baseline SMT system instead of Lucy.
- **Unknown words by SMT**: Since Lucy is flagging unknown words, we translated these individually with the baseline SMT system.

Finally, we experimented with normalization of the punctuation (which was previously included in the pre-processing steps of SMT but not in RBMT), addition of quotes on the menu items and some additional automatic source pre-processing in order to remove redundant phrases such as "where it says".

We ran exhaustive search with all possible combinations of the modification above and the most indicative automatic scores are shown in table 3. Although automatic scores have in the past shown low performance when evaluating RBMT systems, our proposed modifications have a lexical impact that can be adequately measured with n-gram based metrics. Our investigation and discussion is performed on Batch 2. The best combination of the suggested modifications achieves an overall improvement of 0.51 points BLEU and 0.68 points METEOR over the baseline. In particular:

- Adding quotes around menu items resulted in a significant drop of the automatic scores, so it was not used; this needs to be further evaluated, as references do not use quotes for menu items either. Nevertheless, quotes were not always useful due to an occasional erroneous identification of menu item boundaries.
- Separate translation of the menu items (sepMenus) gives a positive result of about 0.46 BLEU and 0.63 METEOR.
- Normalizing punctuation (normPunct) has a slightly positive effect when the menu items are translated separately by Lucy.
- Passing only RBMT's unknown words (unk) to SMT results in a loss of 0.4 BLEU.
- Translating the RBMT's menus with SMT (SMTmenus) also deteriorates the scores and
- translating both menu items and unknown words with SMT (unk+SMTmenus) has a positive effect against the baseline and it seems to be comparable with the best system without SMT (sepMenus+normPunct).

The phrase "where it says" appears in 7% of the sentences in Batch 2 and 2% of the sentences in Batch 1. Although the removal of "where it says" on the source sentence seems to slightly lower the

	BLEU	METEOR	manual
baseline	24.90	44.38	
quotes	24.00	44.29	
sepMenus	25.39	45.01	
sepMenus + normPunct	25.41	45.06	15.8%
sepMenus + normPunct - WhereItSays*	25.36	45.00	84.2%
SMTmenus	24.06	42.83	
unk	24.50	44.05	
unk + sepMenus	23.68	43.30	
unk + SMTmenus	25.41	44.95	
unk + SMTmenus - WhereItSays*	25.36	44.88	

Table 3: Improvements on the RBMT system. (*) indicates the submitted variations.

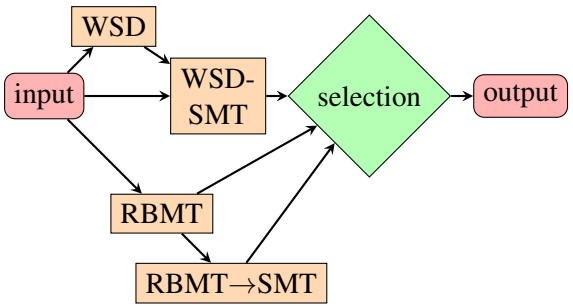

Figure 1: Architecture of the selection mechanism

automatic scores, the difference does not seem significant, and manual inspection raised the concern that this may be because of the way this phrase has been translated in the references. We therefore conducted manual sentence selection on 38 (out of the 69) sentences where this phrase appeared and in 84.2% of the cases its removal made the translation preferable. We therefore concluded in selecting this variation, despite the slightly lower scores.

2.5 Serial RBMT post-editing with SMT

As an alternative to automatic post-editing of the RBMT system, a serial RBMT+SMT system combination is used, as described in (Simard et al., 2007). For building it, the first stage is translation of the source language part of the training corpus by the RBMT system. In the second stage, a SMT system is trained using the RBMT translation output as a source language and the target language part as a target language. Later, the test set is first translated by the RBMT system, and the obtained translation is translated by the SMT system.

2.6 Selection mechanism

The selection mechanism aims to combine various systems, by selecting the best MT output for every sentence. The architecture of the system is illustrated in figure 1. The core of the selection mechanism is a ranker which reproduces ranking by aggregating pairwise decisions by a binary classifier (Avramidis, 2013). Such a classifier is trained on binary comparisons in order to select the best one out of two different MT outputs given one source sentence at a time. As training material, we used the test-sets of WMT evaluation task (2008-2014). The rank labels for the training are automatically generated, after ordering the given MT outputs based on their sentence-level METEOR (Lavie and Agarwal, 2007) against the references. We have previously experimented with training on ranking provided by users, but experiments showed that for this task, ranks made out of sentence-level METEOR maximize all automatic scores on our development set, including other document-level ones, such as BLEU.

We exhaustively tested the available feature sets with many machine learning methods and Support Vector Machines seemed to give the best performance. The binary classifiers were wrapped into rankers using the *soft pairwise recomposition* (Avramidis, 2013) to reduce ties between the systems. Due to technical reasons, the version of the selection mechanism that is submitted to this task is only a pilot version that includes WSD-SMT (section 2.2), baseline RBMT (section 2.4) and RBMT→SMT (section 2.5). When ties occurred, despite the soft recomposition, the system was selected based on a predefined system priority (WSD-SMT, RBMT, RBMT→SMT). The pre-

defined order of the systems needs to be further confirmed as part of the future work.

3 Manual evaluation

Apart from the automatic evaluation scores, we include manual evaluation performed by a professional German linguist.

3.1 Manual evaluation methodology

The manual evaluation was performed in four phases:

- The annotator reads through the development set translated by all systems and identifies the phenomena where often errors occur.
- For each one of the prominent linguistic phenomena, the annotator selects 100 source segments including the respective phenomenon that is prone to MT errors.
- The total occurrences of each phenomenon in all source segments are counted (each phenomenon may occur more than once in a segment, and each segment may contain more than one sentences).
- Consequently, the annotator counts the times each phenomenon has been translated correctly. For a translation to be correct it does not have to be identical with the reference translation. This is repeated for the output of every MT system. The accuracy is calculated as the ratio of the correct translations of the phenomenon divided by the occurrences of the phenomenon in the source.

3.2 Manual evaluation results

The most prominent error categories were found to be **imperatives, compounds, quotation marks, menu item sequences** (separated by ">"), **missing verbs, phrasal verbs** and **terminology**. In these 7 categories, 657 source segments were chosen from development set Batch 2 to demonstrate the phenomena bound to the frequent errors[1]. Many segments contained multiple instances of the respective phenomena, resulting in 2104 instances of phenomena in overall. The results appear in table 4.

The two baseline systems **SMT** and **RBMT** seem to have complementary behavior regarding the investigated phenomena. SMT performs well on terminology, menu items and quotation marks, but seems to suffer on imperatives, missing verbs, phrasal verbs and generation of compounds. On the contrary, RBMT does relatively well with imperatives, compounds, verbs and phrasal verbs, whereas it has issues with menu items and is relatively worse with terminology.

The linear combination system **RBMT→SMT** manages to successfully combine the performance of the two systems regarding imperatives and maintains almost the same performance on verbs and terminology, whereas all other phenomena deteriorate, despite achieving higher automatic scores in overall.

The **SMT-syntax** and the **SMT-WSD** systems seem to have relatively lower performance in all categories.[2] Since the performance of the WSD analyzer has already been confirmed, the failure of the SMT-WSD system to achieve a good performance on terminology and high n-gram-based automatic scores may be an indication that the current data setting does not face ambiguity issues and the senses probably only add additional complexity.

The **selection mechanism** (which in its current version only included SMT-WSD, RBMT and RBMT→SMT) performs better with the terminology and the quotation marks, whereas it maintains the good performance of its components on verbs and menu items. Performance on phrasal verbs nevertheless suffers. Additionally it achieves the highest accuracy on the selected phenomena, with 2% less errors than its best component, the baseline RBMT system.

The two improved versions of the RBMT system appear to have solved the problems they were developed for, namely the compounded menu items and one of them also does better with the quotation marks. The performance on imperatives, verbs and terminology remains the same, but the deterioration on phrasal verbs is obvious. A post-mortem analysis attributes this loss to a logical bug in the menu items detection, which often erroneously included title-cased verbs in the beginning of the sentence, preventing them from being translated as an active part of the sentence.

[1] Despite the goal of collecting 100 segments per category, it was possible to find only 57 segments with phrasal verbs within the development set Batch 2.

[2] A pre-processing bug prevented SMT-syntax from translating quotation marks.

	#	SMT	SMT-WSD	SMT-syntax	RBMT	RBMT →SMT	RBMT menus	RBMT SMTm	sel mech
imperatives	247	68%	65%	68%	**79%**	**83%**	**79%**	**79%**	77%
compounds	220	55%	41%	56%	87%	64%	**89%**	**86%**	78%
"$>$" separators	148	**99%**	75%	97%	39%	66%	84%	80%	74%
quotation marks	431	97%	93%	0%	94%	86%	75%	95%	**98%**
verbs	504	85%	73%	81%	**93%**	**92%**	**93%**	**93%**	**92%**
phrasal verbs	89	22%	3%	7%	**69%**	51%	29%	29%	24%
terminology	465	**64%**	52%	52%	50%	**62%**	54%	53%	**60%**
average		76%	65%	52%	77%	77%	75%	**78%**	**79%**

Table 4: Translation accuracy on manually evaluated sentences focusing on particular phenomena. Test-sets consist of hand-picked source sentences of Batch 2 that include the respective phenomenon. Bold-face indicates best systems on each phenomenon (row) with a 0.95 confidence level.

4 Discussion and further work

In our shared task submission we included:

(i) the SMT and RBMT baseline systems,

(ii) the syntax-enhanced system (DFKI-syntax),

(iii) the RBMT system with separate menu items, normalization of punctuation and removal of "where it says" (previously appearing as sepMenus+normPunct-WhereItSays, submitted as qtl-RBMT-menus),

(iv) the RBMT system with removal of "where it says", passing menu items and unknown words to SMT (previously appearing as unk+SMTmenus-WhereItSays, submitted as qtl-RBMT-SMTmenus) and

(v) the selection mechanism which includes the systems SMT-WSD, RBMT and RBMT→SMT.

The results of the official evaluation campaign for our systems appear in the table 5. RBMT-menus appears to be slightly better than all the other systems we developed, but the difference with the other RBMT systems is not statistically significant. Nevertheless, it is our only system that competes with another competitor system for the 2nd position. Additionally, it is worth noting the failure of BLEU to correlate with the human preferences, mainly for the systems that relate to RBMT, inline with past observations (Callison-Burch et al., 2006).

In future work, we intend to continue this line of development by including all the individual components in the selection mechanism. Additionally, we would focus on solving issues on the particular phenomena, by employing specialized methods. Finally, we should perform a more in-depth evaluation of the selection mechanism and study how the insights gained from the manual inspection of errors can be translated into features that improve the selection.

	rank	TrueSkill	BLEU
RBMT-SMTmenus	2-6	-0.062	25.4
RBMT baseline	3-6	-0.093	25.2
RMBT-menus	3-6	-0.098	25.2
SMT-syntax	7-8	-0.190	34.8
selection	9	-0.382	29.0
SMT baseline	10	-0.485	34.0

Table 5: Human ranks and automatic scores of our submitted systems on the tests, as a result of the official evaluation. Ranks are given in a range in order to account for confidence intervals.

Acknowledgments

This work has received support from (a) the ECs FP7 (FP7/2007-2013) under grant agreement number 610516: "QTLeap: Quality Translation by Deep Language Engineering Approaches" and (b) the German Federal Ministry of Education and Research (BMBF), Unternehmen Region, instrument Wachstumskern-Potenzial number 03WKP45: "DKT: Digitale Kuratierungstechnologien."

References

Alonso, J. A. and Thurmair, G. (2003). The Comprendium Translator system. In *Proceedings of the Ninth Machine Translation Summit*. International Association for Machine Translation (IAMT).

Avramidis, E. (2013). Sentence-level ranking with quality estimation. *Machine Translation*, 27(Special issue on Quality Estimation):239–256.

Avramidis, E., Popovic, M., and Burchardt, A. (2015). DFKI's experimental hybrid MT system for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation. Workshop on Statistical Machine Translation (WMT-2015), 10th, September 17-18, Lisbon, Portugal*, pages 66–73. Association for Computational Linguistics.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy.

Heafield, K. (2011). KenLM : Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, number 2009, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P. (2010). An Experimental Management System. *The Prague Bulletin of Mathematical Linguistics*, 94(-1):87–96.

Koehn, P., Arun, A., and Hoang, H. (2008). Towards better Machine Translation Quality for the German-English Language Pairs. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 139–142, Columbus, Ohio. Association for Computational Linguistics.

Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of the 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York. Association for Computational Linguistics.

Popović, M., Avramidis, E., Burchardt, A., Hunsicker, S., Schmeier, S., Tscherwinka, C., Vilar, D., and Uszkoreit, H. (2014). Involving language professionals in the evaluation of machine translation. *Language Resources and Evaluation*, 48(4):541–559.

Schwenk, H. and Koehn, P. (2008). Large and Diverse Language Models for Statistical Machine Translation. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-based Post-editing. *Proceedings of the 2007 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (April):508–515.

Srivastava, A., Penkale, S., Groves, D., and Tinsley, J. (2009). Evaluating Syntax-Driven Approaches to Phrase Extraction for MT. In *Proceedings of the 3rd International Workshop on Example-based Machine Translation*, pages 19–28, Dublin, Ireland.

Srivastava, A. K. and Way, A. (2009). Using Percolated Dependencies for Phrase Extraction in SMT. In *Proceedings of the Machine Translation Summit XII*, pages 316–323, Ottawa, Canada.

Stolcke, A. (2002). SRILM an Extensible Language Modeling Toolkit. In *System*, volume 2, pages 901–904. ISCA.

Weissenborn, D., Hennig, L., Xu, F., and Uszkoreit, H. (2015). Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 596–605, Beijing, China. Association for Computer Linguistics.

Zhechev, V. (2009). Unsupervised Generation of Parallel Treebank through Sub-Tree Alignment. *Prague Bulletin of Mathematical Linguistics, Special Issue: Open Source Tools for MT*, \bf 91:89–98.