Munich-Edinburgh-Stuttgart Submissions of OSM Systems at WMT13

Nadir Durrani[1], **Helmut Schmid**[2], **Alexander Fraser**[2],
Hassan Sajjad[3], **Richárd Farkas**[4]
[1]University of Edinburgh – dnadir@inf.ed.ac.uk
[2]Ludwig Maximilian University Munich – schmid,fraser@cis.uni-muenchen.de
[3]Qatar Computing Research Institute – hsajjad@qf.org.qa
[4]University of Szeged – rfarkas@inf.u-szeged.hu

Abstract

This paper describes Munich-Edinburgh-Stuttgart's submissions to the Eighth Workshop on Statistical Machine Translation. We report results of the translation tasks from German, Spanish, Czech and Russian into English and from English to German, Spanish, Czech, French and Russian. The systems described in this paper use OSM (Operation Sequence Model). We explain different pre-/post-processing steps that we carried out for different language pairs. For German-English we used constituent parsing for reordering and compound splitting as preprocessing steps. For Russian-English we transliterated the unknown words. The transliteration system is learned with the help of an unsupervised transliteration mining algorithm.

1 Introduction

In this paper we describe Munich-Edinburgh-Stuttgart's[1] joint submissions to the Eighth Workshop on Statistical Machine Translation. We use our in-house OSM decoder which is based on the operation sequence N-gram model (Durrani et al., 2011). The N-gram-based SMT framework (Mariño et al., 2006) memorizes Markov chains over sequences of minimal translation units (MTUs or tuples) composed of bilingual translation units. The OSM model integrates reordering operations within the tuple sequences to form a heterogeneous mixture of lexical translation and reordering operations and learns a Markov model over a sequence of operations.

Our decoder uses the beam search algorithm in a stack-based decoder like most sequence-based SMT frameworks. Although the model is based on minimal translation units, we use phrases during search because they improve the search accuracy of our system. The earlier decoder (Durrani et al., 2011) was based on minimal units. But we recently showed that using phrases during search gives better coverage of translation, better future cost estimation and lesser search errors (Durrani et al., 2013a) than MTU-based decoding. We have therefore shifted to phrase-based search on top of the OSM model.

This paper is organized as follows. Section 2 gives a short description of the model and search as used in the OSM decoder. In Section 3 we give a description of the POS-based operation sequence model that we test for our German-English and English-German experiments. Section 4 describes our processing of the German and English data for German-English and English-German experiments. In Section 5 we describe the unsupervised transliteration mining that has been done for the Russian-English and English-Russian experiments. In Section 6 we describe the sub-sampling technique that we have used for several language pairs. In Section 7 we describe the experimental setup followed by the results. Finally we summarize the paper in Section 8.

2 System Description

2.1 Model

Our systems are based on the OSM (Operation Sequence Model) that simultaneously learns translation and reordering by representing a bilingual

[1]Qatar Computing Research Institute and University of Szeged were partnered for RU-EN and DE-EN language pairs respectively.

Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 122–127,
Sofia, Bulgaria, August 8-9, 2013

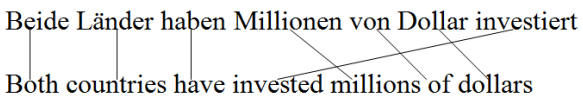

Figure 1: Bilingual Sentence with Alignments

sentence pair and its alignments as a unique sequence of operations. An operation either jointly generates source and target words, or it performs reordering by inserting gaps or jumping to gaps. We then learn a Markov model over a sequence of operations $o_1, o_2, \ldots, o_J$ that encapsulate MTUs and reordering information as:

$$p_{osm}(o_1, ..., o_J) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, ..., o_{j-1})$$

By coupling reordering with lexical generation, each (translation or reordering) decision depends on $n-1$ previous (translation and reordering) decisions spanning across phrasal boundaries. The reordering decisions therefore influence lexical selection and vice versa. A heterogeneous mixture of translation and reordering operations enables us to memorize reordering patterns and lexicalized triggers unlike the classic N-gram model where translation and reordering are modeled separately.

2.2 Training

During training, each bilingual sentence pair is deterministically converted to a unique sequence of operations.[2] The example in Figure 1(a) is converted to the following sequence of operations:

Generate(Beide, Both) → *Generate(Länder, countries)* → *Generate(haben, have)* → *Insert Gap* → *Generate(investiert, invested)*

At this point, the (partial) German and English sentences look as follows:

Beide Länder haben ☐ investiert

Both countries have invested

The translator then jumps back and covers the skipped German words through the following sequence of operations:

Jump Back(1) → *Generate(Millionen, millions)* → *Generate(von, of)* → *Generate(Dollar, dollars)*

[2]Please refer to Durrani et al. (2011) for a list of operations and the conversion algorithm.

The generative story of the OSM model also supports discontinuous source-side cepts and source-word deletion. However, it doesn't provide a mechanism to deal with unaligned and discontinuous target cepts. These are handled through a 3-step process[3] in which we modify the alignments to remove discontinuous and unaligned target MTUs. Please see Durrani et al. (2011) for details. After modifying the alignments, we convert each bilingual sentence pair and its alignments into a sequence of operations as described above and learn an OSM model. To this end, a Kneser-Ney (Kneser and Ney, 1995) smoothed 9-gram model is trained with SRILM (Stolcke, 2002) while KenLM (Heafield, 2011) is used at runtime.

2.3 Feature Functions

We use additional features for our model and employ the standard log-linear approach (Och and Ney, 2004) to combine and tune them. We search for a target string E which maximizes a linear combination of feature functions:

$$\hat{E} = \arg\max_{E} \left\{ \sum_{j=1}^{J} \lambda_j h_j(o_1, ..., o_J) \right\}$$

where λ_j is the weight associated with the feature $h_j(o_1, ..., o_j)$. Apart from the main OSM feature we train 9 additional features: A target-language model (see Section 7 for details), 2 lexical weighting features, gap and open gap penalty features, two distance-based distortion models and 2 length-based penalty features. Please refer to Durrani et al. (2011) for details.

2.4 Phrase Extraction

Phrases are extracted in the following way: The aligned training corpus is first converted to an operation sequence. Each subsequence of operations that starts and ends with a translation operation, is considered a "phrase". The translation operations include *Generate Source Only (X)* operation which deletes unaligned source word. Such phrases may be discontinuous if they include reordering operations. We replace each subsequence of reordering operations by a discontinuity marker.

[3]Durrani et al. (2013b) recently showed that our post-processing of alignments hurt the performance of the Moses Phrase-based system in several language pairs. The solution they proposed has not been incorporated into the current OSM decoder yet.

During decoding, we match the source tokens of the phrase with the input. Whenever there is a discontinuity in the phrase, the next source token can be matched at any position of the input string. If there is no discontinuity marker, the next source token in the phrase must be to the right of the previous one. Finally we compute the number of uncovered input tokens within the source span of the hypothesized phrase and reject the phrase if the number is above a threshold. We use a threshold value of 2 which had worked well in initial experiments. Once the positions of all the source words of a phrase are known, we can compute the necessary reordering operations (which may be different from the ones that appeared in the training corpus). This usage of phrases allows the decoder to generalize from a seen translation "scored a goal – ein Tor schoss" (where scored/a/goal and schoss/ein/Tor are aligned, respectively) to "scored a goal – schoss ein Tor". The phrase can even be used to translate "er schoss heute ein Tor – he scored a goal today" although "heute" appears within the source span of the phrase "ein Tor schoss". Without phrase-based decoding, the unusual word translations "schoss–scored" and "Tor–goal" (at least outside of the soccer literature) are likely to be pruned.

The phrase tables are further filtered with threshold pruning. The translation options with a frequency less than x times the frequency of the most frequent translation are deleted. We use $x = 0.02$. We use additional settings to increase this threshold for longer phrases. The phrase filtering heuristic was used to speed up decoding. It did not lower the BLEU score in our small scale experiments (Durrani et al., 2013a), however we could not test whether this result holds in a large scale evaluation.

2.5 Decoder

The decoding framework used in the operation sequence model is based on Pharaoh (Koehn, 2004). The decoder uses beam search to build up the translation from left to right. The hypotheses are arranged in m stacks such that stack i maintains hypotheses that have already translated i many foreign words. The ultimate goal is to find the best scoring hypothesis, that translates all the words in the foreign sentence. During the hypothesis extension each extracted phrase is translated into a sequence of operations. The reordering operations (gaps and jumps) are generated by looking at the position of the translator, the last foreign word generated etc. (Please refer to Algorithm 1 in Durrani et al. (2011)). The probability of an operation depends on the $n-1$ previous operations. The model is smoothed with Kneser-Ney smoothing.

3 POS-based OSM Model

Part-of-speech information is often relevant for translation. The word "stores" e.g. should be translated to "Läden" if it is a noun and to "speichert" when it is a verb. The sentence "The small child cries" might be incorrectly translated to "Die kleinen Kind weint" where the first three words lack number, gender and case agreement.

In order to better learn such constraints which are best expressed in terms of part of speech, we add another OSM model as a new feature to the log-linear model of our decoder, which is identical to the regular OSM except that all the words have been replaced by their POS tags. The input of the decoder consists of the input sentence with automatically assigned part-of-speech tags. The source and target part of the training data are also automatically tagged and phrases with words and POS tags on both sides are extracted. The POS-based OSM model is only used in the German-to-English and English-to-German experiments.[4] So far, we only used coarse POS tags without gender and case information.

4 Constituent Parse Reordering

Our German-to-English system used constituent parses for pre-ordering of the input. We parsed all of the parallel German to English data available, and the tuning, test and blind-test sets. We then applied reordering rules to these parses. We used the rules for reordering German constituent parses of Collins et al. (2005) together with the additional rules described by Fraser (2009). These are applied as a preprocess to all German data (training, tuning and test data). To produce the parses, we started with the generative BitPar parser trained on the Tiger treebank with optimizations of the grammar, as described by (Fraser et al., 2013). We then performed self-training using the high quality Europarl corpus - we parsed it, and then retrained the parser on the output.

[4]This work is ongoing and we will present detailed experiments in the future.

Following this, we performed linguistically-informed compound splitting, using the system of Fritzinger and Fraser (2010), which disambiguates competing analyses from the high-recall Stuttgart Morphological Analyzer SMOR (Schmid et al., 2004) using corpus statistics (Koehn and Knight, 2003). We also split portmanteaus like German "zum" formed from "zu dem" meaning "to the". Due to time constraints, we did not address German inflection. See Weller et al. (2013) for further details of the linguistic processing involved in our German-to-English system.

5 Transliteration Mining/Handling OOVs

The machine translation system fails to translate out-of-vocabulary words (OOVs) as they are unknown to the training data. Most of the OOVs are named entities and simply passing them to the output often produces correct translations if source and target language use the same script. If the scripts are different transliterating them to the target language script could solve this problem. However, building a transliteration system requires a list of transliteration pairs for training. We do not have such a list and making one is a cumbersome process. Instead, we use the unsupervised transliteration mining system of Sajjad et al. (2012) that takes a list of word pairs for training and extracts transliteration pairs that can be used for the training of the transliteration system. The procedure of mining transliteration pairs and transliterating OOVs is described as follows:

We word-align the parallel corpus using GIZA++ in both direction and symmetrize the alignments using the grow-diag-final-and heuristic. We extract all word pairs which occur as 1-to-1 alignments (like Sajjad et al. (2011)) and later refer to them as the *list of word pairs*. We train the unsupervised transliteration mining system on the list of word pairs and extract transliteration pairs. We use these mined pairs to build a transliteration system using the Moses toolkit. The transliteration system is applied in a post-processing step to transliterate OOVs. Please refer to Sajjad et al. (2013) for further details on our transliteration work.

6 Sub-sampling

Because of scalability problems we were not able to use the entire data made available for building the translation model in some cases. We used modified Moore-Lewis sampling (Axelrod et al., 2011) for the language pairs es-en, en-es, en-fr, and en-cs. In each case we included the News-Commentary and Europarl corpora in their entirety, and scored the sentences in the remaining corpora (the selection corpus) using a filtering criterion, adding 10% of the selection corpus to the training data. We can not say with certainty whether using the entire data will produce better results with the OSM decoder. However, we know that the same data used with the state-of-the-art Moses produced worse results in some cases. The experiments in Durrani et al. (2013c) showed that MML filtering decreases the BLEU scores in es-en (news-test13: Table 19) and en-cs (news-test12: Table 14). We can therefore speculate that being able to use all of the data may improve our results somewhat.

7 Experiments

Parallel Corpus: The amount of bitext used for the estimation of the translation models is: de–en $\approx$ 4.5M and ru–en $\approx$ 2M parallel sentences. We were able to use all the available data for cs-to-en ($\approx$ 15.6M sentences). However, sub-sampled data was used for en-to-cs ($\approx$ 3M sentences), en-to-fr ($\approx$ 7.8M sentences) and es–en ($\approx$ 3M sentences).

Monolingual Language Model: We used all the available training data (including LDC Gigaword data) for the estimation of monolingual language models: en $\approx$ 287.3M sentences, fr $\approx$ 91M, es $\approx$ 65.7M, cs $\approx$ 43.4M and ru $\approx$ 21.7M sentences. All data except for ru-en and en-ru was true-cased. We followed the approach of Schwenk and Koehn (2008) by training language models from each sub-corpus separately and then linearly interpolated them using SRILM with weights optimized on the held-out dev-set. We concatenated the news-test sets from four years (2008-2011) to obtain a large dev-set[5] in order to obtain more stable weights (Koehn and Haddow, 2012).

Decoder Settings: For each extracted input phrase only 15-best translation options were used during decoding.[6] We used a hard reordering limit

[5]For Russian-English and English-Russian language pairs, we divided the tuning-set news-test 2012 into two halves and used the first half for tuning and second for test.

[6]We could not experiment with higher n-best translation options due to a bug that was not fixed in time and hindered us from scaling.

of 16 words which disallows a jump beyond 16 source words. A stack size of 100 was used during tuning and 200 for decoding the test set.

Results: Table 1 shows the uncased BLEU scores along with the rank obtained on the submission matrix.[7] We also show the results from human evaluation.

Lang	Evaluation			
	Automatic		Human	
	BLEU	Rank	Win Ratio	Rank
de-en	27.6	9/31	0.562	6-8
es-en	30.4	6/12	0.569	3-5
cs-en	26.4	3/11	0.581	2-3
ru-en	24.5	8/22	0.534	7-9
en-de	20.0	6/18		
en-es	29.5	3/13	0.544	5-6
en-cs	17.6	14/22	0.517	4-6
en-ru	18.1	6/15	0.456	9-10
en-fr	30.0	7/26	0.541	5-9

Table 1: Translating into and from English

8 Conclusion

In this paper, we described our submissions to WMT 13 in all the shared-task language pairs (except for fr-en). We used an OSM-decoder, which implements a model on n-gram of operations encapsulating lexical generation and reordering. For German-to-English we used constituent parsing and applied linguistically motivated rules to these parses, followed by compound splitting. We additionally used a POS-based OSM model for German-to-English and English-to-German experiments. For Russian-English language pairs we used unsupervised transliteration mining. Because of scalability issues we could not use the entire data in some language pairs and used only sub-sampled data. Our Czech-to-English system that was built from the entire data did better in both automatic and human evaluation compared to the systems that used sub-sampled data.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful feedback and suggestions. We would like to thank Philipp Koehn and Barry Haddow for providing data and alignments. Nadir Durrani was funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658. Alexander Fraser was funded by Deutsche Forschungsgemeinschaft grant Models of Morphosyntax for Statistical Machine Translation. Helmut Schmid was supported by Deutsche Forschungsgemeinschaft grant SFB 732. Richárd Farkas was partially funded by the Hungarian National Excellence Program (TÁMOP 4.2.4.A/2-11-1-2012-0001). This publication only reflects the authors' views.

[7] http://matrix.statmt.org/

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *ACL05*, pages 531–540, Ann Arbor, MI.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A Joint Sequence Translation Model with Integrated Reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June.

Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013a. Model With Minimal Translation Units, But Decode With Phrases. In *The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013b. Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nadir Durrani, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013c. Edinburgh's Machine Translation Systems for European Language Pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alexander Fraser, Helmut Schmid, Richárd Farkas, Renjing Wang, and Hinrich Schütze. 2013. Knowledge sources for constituent parsing of German, a morphologically rich and less-configurational language. *Computational Linguistics - to appear*.

Alexander Fraser. 2009. Experiments in Morphosyntactic Processing for Translating to and from German. In *Proceedings of the EACL 2009 Fourth Workshop on Statistical Machine Translation*, pages 115–119, Athens, Greece, March.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the ACL 2010 Fifth Workshop on Statistical Machine Translation*, Uppsala, Sweden.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, 7.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan, May.

Philipp Koehn and Barry Haddow. 2012. Towards Effective Use of Training Data in Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 317–321, Montréal, Canada, June. Association for Computational Linguistics.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 187–193, Morristown, NJ.

Philipp Koehn. 2004. Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *AMTA*, pages 115–124.

José B. Mariño, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-Based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Franz J. Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(1):417–449.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2011. An algorithm for unsupervised transliteration mining with an application to word alignment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, USA.

Hassan Sajjad, Alexander Fraser, and Helmut Schmid. 2012. A statistical model for unsupervised and semi-supervised transliteration mining. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, Jeju, Korea.

Hassan Sajjad, Svetlana Smekalova, Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013. QCRI-MES Submission at WMT13: Using Transliteration Mining to Improve Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Holger Schwenk and Philipp Koehn. 2008. Large and Diverse Language Models for Statistical Machine Translation. In *International Joint Conference on Natural Language Processing*, pages 661–666, January 2008.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Intl. Conf. Spoken Language Processing*, Denver, Colorado.

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas. 2013. Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August. Association for Computational Linguistics.