Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation

Alexandra Birch, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, Philipp Koehn

School of Informatics
University of Edinburgh
Scotland, United Kingdom

a.birch@ed.ac.uk {mhuck,dnadir,nbogoych,pkoehn}@inf.ed.ac.uk

Abstract

This paper describes the University of Edinburgh's spoken language translation (SLT) and machine translation (MT) systems for the IWSLT 2014 evaluation campaign. In the SLT track, we participated in the German↔English and English→French tasks. In the MT track, we participated in the German↔English, English→French, Arabic↔English, Farsi→English, Hebrew→English, Spanish↔English, and Portuguese-Brazil↔English tasks.

For our SLT submissions, we experimented with comparing operation sequence models with bilingual neural network language models. For our MT submissions, we explored using unsupervised transliteration for languages which have a different script than English, in particular for Arabic, Farsi, and Hebrew. We also investigated syntax-based translation and system combination.

1. Introduction

The University of Edinburgh's translation engines are based on the open source Moses toolkit [1]. We set up phrase-based systems [2] for all SLT and MT tasks covered in this paper, and additionally a string-to-tree syntax-based system [3, 4] for the English→German MT task.

The setups for our phrase-based systems have evolved from the configurations of the engines we built for last year's IWSLT [5] and for this year's Workshop on Statistical Machine Translation (WMT) [6]. The notable features of these systems are:

- Phrase translation scores in both directions, smoothed with Good-Turing discounting
- Lexical translation scores in both directions
- Word and phrase penalties
- Six simple count-based binary features
- Phrase length features
- Distance-based distortion cost
- A hierarchical lexicalized reordering model [7]
- Sparse lexical and domain indicator features [8]
- Operation sequence models (OSMs) over different word representations [9, 10]
- A 5-gram language model (LM) over words

We typically train factored phrase-based translation models [11, 12] and also incorporate higher order *n*-gram LMs over word representations given by the factors. Factors can for instance be lemma, part-of-speech (POS) tag, morphological tag, or automatically learnt word classes in the manner of Brown clusters [13].

Edinburgh's syntax-based systems have recently yielded state-of-the-art performance on English→German news translation tasks [14, 15] but have not been applied in an IWSLT-style setting before. Standard features of our string-to-tree syntax-based systems are:

- Rule translation scores in both directions, smoothed with Good-Turing discounting
- Lexical translation scores in both directions
- Word and rule penalties
- A rule rareness penalty
- The monolingual PCFG probability of the tree fragment from which the rule was extracted
- A 5-gram LM over words

For our Spanish↔English and Portuguese-Brazil↔English submissions, we ran the engines as described in last year's system description paper [5]. In the following, we focus on describing the new systems which were developed for the rest of the tasks.

Our this year's IWSLT systems were trained using monolingual and parallel data from WIT^3 [16], Europarl [17], MultiUN [18], the Gigaword corpora as provided by the Linguistic Data Consortium [19], the German Political Speeches Corpus [20], and the corpora provided for the WMT shared translation task [21].

Word alignments for the MT track systems were created by aligning the data in both directions with MGIZA++ [22] and symmetrizing the two alignments with the grow-diag-final-and heuristic [23, 2]. Word alignments for the SLT track systems were created using fast_align [24].

The SRILM toolkit [25] was employed to train 5-gram language models (LMs) with modified Kneser-Ney smoothing [26]. We trained individual LMs on each corpus and then interpolated them using weights tuned to minimize perplexity on a development set. KenLM [27] was utilized for LM scoring during decoding. Model weights for the log-linear model combination [28] were optimized with batch *k*-best MIRA [29] to maximize BLEU [30]. Where not otherwise stated, the systems were tuned on dev2010.

Besides participating in the evaluation campaign with our individual engines, we also collaborated with partners from the EU-BRIDGE project to produce additional joint submissions. The combined systems of the University of Edinburgh, RWTH Aachen University, Karlsruhe Institute of Technology, and Fondazione Bruno Kessler are described in [31].

2. Spoken Language Translation

Edinburgh's spoken language translation system experiments set out to compare two recent strands of research in terms of their performance and their properties in order to understand the contributions of each. The first strand of research is bilingual neural network langauge models. There has recently been a great deal of interest bilingual neural network language models as they have shown strong gains in performance for Arabic→English, and to a lesser extent for Chinese→English [32]. It is still not clear what the exact contribution of the bilinugal language model is, and there is reason to believe that its contribution may be that it allows the SMT model to overcome strong phrase pair independence assumptions.

The second strand of research is operation sequence modelling [33, 34]. The integration of the OSM model into phrase-based decoding directly addresses the problem of the phrasal independence assumption by modelling the context of phrase pair translations. We aim to compare these two different approaches and combining them. As we see, combining OSM and the bilingual NN language model slightly outperforms all other models, including the state-of-the-art OSM model, but only for English→French and only very slightly.

2.1. Baseline

For the SLT track, we trained phrase-based models using Moses with mostly default settings. We further included basic sparse features [35] and we used factors. For German→English we used POS tags, morphological tags and lemmas as factors in decoding [11], and for English→German we used POS tags and morphological tags on the target side. Table 1 lists the factors used for the translation model, and the factors over which we trained OSM models.

The SLT and the MT systems were trained in a similar fashion, with the main difference being that for SLT no pre-reordering was performed for German→English as this relies on grammatically correct test sentences, and automatic speech recognition (ASR) output, especially for German, is difficult to parse correctly. We trained the SLT systems on the Europarl, WIT^3, News Commentary, and Commoncrawl corpora. The monolingual data contained the target side of the parallel corpora, the news language model data provided for WMT, and the LDC Gigaword for French and English. The number of words of training data can be seen in Table 1.

	EN→FR	EN→DE	DE→EN
Src Factors	w,c	w,c	w,l,p,m
Tgt Factors	w,c	w,p,m,c	w,l,p
OSM	w,c	w,c	w,l,p and m→p
No. words ∥	138M/153M	116M/110M	110M/116M
No. words mono	2673M	2214M	6600M

Table 1: SLT track: Factors used by translation models and OSM models (words w, clusters c, lemma l, pos p, morphology m) and the size of the parallel and monolingual training data in millions of words.

2.2. Monolingual Punctuation Models

One of the main challenges of spoken language translation is to overcome the mismatch in the style of data that the speech recognition systems output, and the written text that is used to train the translation model. ASR system output lacks punctuation and capitalisation and this is one of the main stylistic differences. Previous research [36, 5] suggests that it is preferrable to punctuate the text before translation, which is what we did by training a monolingual translation system for our two source languages: German and English. The "source language" of the punctuation model has punctuation and capitalisation stripped, and the "target language" is the full original written text. Our handling of punctuation uses a phrase-based translation model with no distortion or reordering, and we tuned the model to the ASR input text (dev2010 for English, and dev2012 for German) using batch MIRA and the BLEU score. After running ASR output through the punctuation model, it is then translated with a standard machine translation model, trained directly on the parallel written text, in a very similar fashion to the MT system, except that for our official submission we tuned the MT model to the ASR tuning set.

2.3. Operation Sequence Model

We investigated applying a number of OSM models [33, 34] to the basic phrase-based translation model. OSM addresses the problem of the phrasal independence assumption since the model considers context beyond phrasal boundaries. The OSM model represents a bilingual sentence pair and its alignment through a sequence of operations that generate the aligned sentence pair. An operation either generates source and target words or it performs reordering by inserting gaps and jumping forward and backward. It has shown to improve performance over many language pairs, and to help even more when sequence models are applied over more general factors such as POS tags and GIZA++'s `mkcls` clusters [5]. For this experiment we applied the best OSM settings from last year's IWSLT experiments which included models over words, lemmas, POS tags, and clusters depending on the language pair. See Table 1 for details.

2.4. Bilingual Neural Network Language Model

There has recently been a great deal of interest in including neural networks in machine translation [37, 38]. There is hope that neural networks provide a way to relax some of the more egregious independence assuptions made in translation models. The challenge with neural networks however, is that they are computationally very expensive, and getting them to operate at scale requires sophisticated efficiency techniques. A recent paper which was able to fully integrate a neural network which includes both source side and target side context in decoding [32], and they managed to show big improvements for a small Arabic→English task, and smaller improvements for a Chinese→English task. We implemented a bilingual neural network language model in order to investigate what their benefits are to state-of-the-art translation models.

We implemented a BiNNLM as a feature function inside Moses, following closely the implementation outlined in [32]. The main focus of our design is to make the Moses specific code flexible and independent of the neural network language model that would be used for scoring. As a result any NNLM could implement the interface and be used by Moses during decoding. Some features such as backoff to POS tag in case of unknown word or use of special $< null >$ token to pad an incomplete parse in the chart decoder are made optional. Currently the implemented backends are NPLM [39] and OxLM [40]. Implementation is available for both phrase based and hierarchical Moses. For our experiments we chose NPLM to be our NNLM backend. We chose it, because it features noise contrastive estimation (NCE) which allows us to avoid having to apply softmax to normalize the outputs, as it is infeasible to do so with large vocabularies. Another benefit of NPLM is that when using NCE and a neural network with one hidden layer we can precompute the values for the first hidden layer of all vocabulary terms, similarly to what [32] do. We also modified the NPLM code a bit and used Magma enabled fork of the Eigen library[1] to speed up the training. This results in a decoder which is about twice as slow as the phrase-based decoder without BiNNLM On average decoding speed is three sentences per second when using BiNNLM, which highlights that this implementation is fast enough to make large experiments possible.

For these experiments we used a target context of four words, and an aligned source window of nine words. Note that NPLM does not support separate source and target contexts so what we did is use the parallel corpora to extract 14-grams which consist of 9 source and 5 target words. Once those 14-grams are extracted we train NPLM on them as if it were a monolingual dataset. The size of our word embedding layers was 256 for the EN→FR, and 150 for DE→EN language models. Increasing the size of the embeddings for DE→EN did not increase performance, but decreasing it for EN→FR seemed to hurt performance. We used just one hidden layer to allow precomputation and much faster decoding. We used a source and target vocabulary size of 16k words, and used a part-of-speech backoff for the less frequent words for the DE→EN system, and backoff to the UNK token for EN→FR.

	EN→FR	DE→EN
Baseline	35.7	32.5
OSM full	37.3	33.0
BiNNLM	36.7	32.4
OSM + BiNNLM	37.4	32.8

Table 2: Performance comparison of OSM and BiNNLM (average case-sensitive BLEU score of IWSLT test sets 2010-2012).

	EN→FR	DE→EN	EN→DE
dev2012	-	21.00	-
dev2010	23.39	-	21.25
test2014	25.50	17.67	17.00

Table 3: Results of submission systems in the SLT track (case-sensitive BLEU scores).

2.5. Results

Looking at Table 2 it seems that both the OSM model and the BiNNLM model outperform the baseline. The OSM model is stronger than the BiNNLM when both features are used separately. However, for the EN→FR task, combining OSM and BiNNLM outperforms OSM on its own by 0.14 BLEU points. The baseline translation systems use large amounts of parallel and monolingual data, and it is not surprising that our first attempt at using BiNNLM did not resoundingly beat the previously state-of-the-art OSM models. It is surprising perhaps that BiNNLM did much better for EN→FR than DE→EN. This is similar to the Devlin et al. result where their AR→EN improvements were much stronger than their ZH→EN results.

From the results here it does seem like the advantages gained by applying OSM and BiNNLM might overlap, given that there is not a large improvement seen when combining the two types of features.

We used the baseline systems trained with OSM models for our official submission to the IWSLT 2014 evaluation. We tuned these on the supplied ASR development sets. The results are shown in Table 3.

3. Machine Translation

This section contains a description of the experiments we carried out for tasks in the MT track of the evaluation campaign.

[1] https://github.com/bravegag/eigenmagma

Pair	Training	**tst2010**	**tst2011**	**tst2012**
	-	26.7	26.3	29.8
AR→EN	7.6K	26.8	26.5	29.9
OOV		393	345	442
	-	8.8	9.6	9.5
EN→AR	9.1K	8.8	9.7	9.6
OOV		351	277	424
	-	15.6	20.7	15.6
FA→EN	5.5K	15.8	21.0	15.8
OOV		337	451	628
	-	30.1	31.5	31.7
HE→EN	14K	30.3	31.8	31.9
OOV		837	753	892

Table 4: Effect of unsupervised transliteration models. Training = extracted transliteration corpus (types). First rows: system without transliteration. Second rows: transliterating OOVs. Third rows: number of OOVs (types) in each test.

EN→AR	**tst2010**	**tst2011**	**tst2012**
baseline	8.3	8.3	8.7
+ Gigaword + UN	8.9	9.2	9.6

Table 5: Effect of Gigaword and UN monolingual data on English→Arabic translation quality.

3.1. Unsupervised Transliteration Model

Arabic, Farsi and Hebrew are written in different writing scripts as English, therefore the conventional method of copying unknown words to the output is not a good idea. We built unsupervised transliteration models [41] to translate OOV words.

The transliteration model is induced using an EM-based method [42]. We extracted transliteration pairs automatically from the word-aligned parallel data and used it to learn a transliteration system. We then built transliteration phrase-tables for translating OOV words and used the post-decoding method (Method 2 as described in the paper) to translate these. Table 4 show results from using unsupervised transliteration models. Small improvements were shown in all cases. Note that not all the OOVs can be translated correctly through transliteration. Only a handful of these were named entities and foreign words that could be transliterated.

3.2. Arabic-English MT

We carried out a number of experiments for the Arabic-English language pair which we now discuss briefly.

Tokenization. We used MADA tokenizer for source-side Arabic [43] and tried different segmentation schemes including D*, S2 and ATB. The ATB segmentation consistently outperformed other schemes.

Modified Moore and Lewis Filtering. The in-domain datasets (TED talk corpus) are small and a large out-of-domain corpus (UN) is available. We tried to explore various ways to make best use of the out-of-domain data to improve the baseline system. We used Modified Moore and Lewis as known as MML [44] filtering, to subsample training data that is similar to the in-domain data. We varied the percentage of bilingual UN data selected between 2%, 5%, 20% and 100%. Adding any percentage of UN data did not give any gains in the performance. Using 2% gave best results, however, they were still below the baseline system.

Backoff Phrase Tables. Instead of using UN data directly we used it with the *backoff* phrase-table method. This allows Moses to use the phrase-table built with the UN data only when a phrase is unknown to phrase-table trained from the in-domain data. The backoff order determines the maximum phrase length for which this operation is allowed. We used backoff order of 5. Using backoff phrase tables gave slight improvement in English→Arabic, results stayed constant or dropped in Arabic→English direction.

Class-based Model. We explored the use of automatic word clusters in phrase-based models [10]. We computed the clusters with GIZA++'s `mkcls` [45] on the source and target side of the parallel training corpus. Clusters are word classes that are optimized to reduce *n*-gram perplexity. By generating a cluster identifier for each output word, we are able to add an *n*-gram model over these identifiers as an additional scoring function. The inclusion of such an additional factor is trivial given the factored model implementation [11] of Moses. The *n*-gram model is trained in the similar way as the regular language model. The lexically driven OSM model falls back to very small context sizes of two to three operations due to data sparsity. Learning operation sequences over cluster-ids enables us to learn richer translation and reordering patterns that can generalize better in sparse data conditions.

Using class-based models, however, did not give any improvements for Arabic-English tasks. We also trained OSM models over cluster-ids. This result contradicts our findings in last year IWSLT paper [5] where we reported significant gains using class-based models on many European language pairs with English as source language.

Monolingual Arabic Data. Unlike parallel data, adding Gigaword and UN monolingual data in English→Arabic translation task gave significant improvements. The gains are shown in Table 5.

3.3. German→English MT

For the German→English MT task system, pre-reordering [46] and compound splitting [47] were applied to the German source language side in a preprocessing step. A factored translation model was employed. Source side factors are word, lemma, POS tag, and morphological tag. Target side factors are word, lemma, and POS tag. Supplementary to the features listed in Section 6, we incorporated two additional LMs into the German→English MT system: a 7-gram LM over POS tags and a 7-gram LM over lemmas (both trained on WIT^3 only). Model weights were optimized on a concatenation of dev2010 and dev2012. Table 6 contains the results on the three test sets.

3.4. English→French MT

We submitted outputs of three phrase-based systems for the English→French MT task: a *primary* system and two contrastive systems (*contrastive 1* and *contrastive 2*). All available training corpora were utilized, with the exception of the MultiUN corpus and the WMT 10^9 French-English corpus, which we excluded from both the parallel and the LM training data. Our systems comprise Brown clusters with 200 classes as additional factors on source and target side. Supplementary to the features listed in Section 6, we incorporated a 7-gram LM over Brown clusters. Furthermore, a bilingual neural network language model as described in Section 2.4 was integrated into the *primary* and the *contrastive 1* system. The primary system was tuned on tst2012, the contrastive systems were tuned on dev2010.

The characteristics of the setup denoted as *contrastive 1* are thus the same as those of the primary submission. We employed identical configuration parameters and features, the only difference between the two systems is the usage of a different tuning set for the optimization of model weights. The setup denoted as *contrastive 2* is similar to *contrastive 1* but does not comprise the bilingual neural network language model. Experimental results are presented in Table 7.

3.5. English→German MT

For the English→German MT task, we submitted outputs of a phrase-based system (*primary*), a syntax-based system (*contrastive 1*), and a system combination (*contrastive 2*). Table 8 shows their respective performance in terms of BLEU scores.

Phrase-based System. The *primary* system is phrase-based with factored models. Source side factors are word, POS tag, and Brown cluster (2000 classes). Target side factors are word, POS tag, Brown cluster (2000 classes), and morphological tag. The primary system was trained with all corpora. Additional features of the primary system are: a 5-gram LM over Brown clusters, a 7-gram LM over morphological tags, and a 7-gram LM over POS tags. Model weights of the primary system were optimized on a concatenation of dev2010 and dev2012.

We trained a second, smaller phrase-based system on in-domain bitexts only (i.e., we restricted the parallel training data to the WIT^3 corpus). We denote this second phrase-based system as *phrase-based in-domain*. Individual hypotheses from the phrase-based in-domain system have not been submitted for the evaluation; we merely added them as auxiliary inputs to our system combination. Additional features of the phrase-based in-domain system are: a 5-gram LM over Brown clusters and a 7-gram LM over morphological tags (the latter trained on WIT^3 only). Model weights of the phrase-based in-domain system were optimized on dev2010.

Syntax-based System. The *contrastive 1* system is a string-to-tree translation system with similar features as the ones described in [15]. The target-side data was parsed with BitPar [48], and right binarization was applied to the parse trees. The system was adapted to the TED domain by extracting separate rule tables (from the WIT^3 corpus and from the rest of the parallel data) and merging them with a fill-up technique [49]. Augmenting the system with non-syntactic phrases [50] and adding soft source syntactic constraints [51] yielded further improvements. Model weights of the syntax-based system were optimized on a concatenation of dev2010 and dev2012.

System Combination. We combined the outputs of the phrase-based primary system, the auxiliary phrase-based in-domain system, and the string-to-tree syntax-based system with the MT system combination approach implemented in the Jane toolkit [52]. The parameters of the system combination were optimized on tst2012. The consensus translation produced by the system combination (syscom) was submitted as *contrastive 2*.

DE→EN	**tst2010**	**tst2011**	**tst2012**
primary	31.6	37.3	31.7

Table 6: Results for the German→English MT task (case-sensitive BLEU scores).

EN→FR	**tst2010**	**tst2011**	**tst2012**
primary	34.4	41.5	44.9
contrastive 1	33.8	40.3	41.4
contrastive 2	33.6	40.2	41.0

Table 7: Results for the English→French MT task (case-sensitive BLEU scores). The contrastive systems were tuned on dev2010, the primary system was tuned on tst2012. A bilingual neural network language model was integrated into *primary* and *contrastive 1*.

EN→DE	**tst2010**	**tst2011**	**tst2012**
phrase-based (primary)	24.9	27.8	23.4
phrase-based in-domain	24.1	26.7	22.2
syntax-based (contrastive 1)	24.8	26.5	23.1
syscom (contrastive 2)	26.0	27.8	24.5

Table 8: Results for the English→German MT task (case-sensitive BLEU scores). The *contrastive 2* submission is a system combination of three systems which was tuned on tst2012.

4. Summary

The Edinburgh submissions for IWSLT cover many language pairs and research techniques. We have implemented a bilingual neural network language model feature in Moses and have demonstrated that it can lead to state-of-the-art results for English→French. BiNNLM seems less beneficial for German→English, however. Our experiments further confirmed the benefit of using OSM, transliteration and system combination.

5. Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 287658 (EU-BRIDGE) and grant agreement n^o 288487 (MosesCore).

6. References

[1] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. J. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, June 2007, pp. 177–180.

[2] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, May/June 2003, pp. 127–133.

[3] M. Galley, M. Hopkins, K. Knight, and D. Marcu, "What's in a translation rule?" in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Boston, MA, USA, May 2004, pp. 273–280.

[4] P. Williams and P. Koehn, "GHKM Rule Extraction and Scope-3 Parsing in Moses," in *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June 2012, pp. 434—440.

[5] A. Birch, N. Durrani, and P. Koehn, "Edinburgh slt and mt system description for the IWSLT 2013 evaluation," in *Proceedings of the 10th International Workshop on Spoken Language Translation*, Heidelberg, Germany, December 2013, pp. 40—48.

[6] N. Durrani, B. Haddow, P. Koehn, and K. Heafield, "Edinburgh's phrase-based machine translation systems for WMT-14," in *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June 2014, pp. 97–104.

[7] M. Galley and C. D. Manning, "A Simple and Effective Hierarchical Phrase Reordering Model," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Honolulu, HI, USA, Oct. 2008, pp. 847–855.

[8] E. Hasler, B. Haddow, and P. Koehn, "Sparse Lexicalised Features and Topic Adaptation for SMT," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, Dec. 2012, pp. 268–275.

[9] N. Durrani, H. Schmid, and A. Fraser, "A joint sequence translation model with integrated reordering," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, June 2011, pp. 1045–1054.

[10] N. Durrani, P. Koehn, H. Schmid, and A. Fraser, "Investigating the Usefulness of Generalized Word Representations in SMT," in *Proceedings of the 25th Annual Conference on Computational Linguistics (COLING)*, Dublin, Ireland, August 2014, pp. 421—432.

[11] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 868–876.

[12] P. Koehn and B. Haddow, "Interpolated Backoff for Factored Translation Models," in *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, Oct./Nov. 2012.

[13] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes," in *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1999, pp. 71–76.

[14] M. Nadejde, P. Williams, and P. Koehn, "Edinburgh's Syntax-Based Machine Translation Systems," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Sofia, Bulgaria, Aug. 2013, pp. 170–176.

[15] P. Williams, R. Sennrich, M. Nadejde, M. Huck, E. Hasler, and P. Koehn, "Edinburgh's Syntax-Based Systems at WMT 2014," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 207–214.

[16] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the 16^{th} Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[17] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proc. of the MT Summit X*, Phuket, Thailand, Sept. 2005.

[18] A. Eisele and Y. Chen, “MultiUN: A Multilingual Corpus from United Nation Documents,” in *Proceedings of the Seventh conference on International Language Resources and Evaluation*, May 2010, pp. 2868–2872.

[19] Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu.

[20] A. Barbaresi, “German Political Speeches, Corpus and Visualization,” ENS Lyon, Tech. Rep., 2012, 2nd Version. [Online]. Available: http://purl.org/corpus/german-speeches

[21] Shared Translation Task of the ACL 2014 Ninth Workshop on Statistical Machine Translation, http://www.statmt.org/wmt14/translation-task.html.

[22] Q. Gao and S. Vogel, “Parallel Implementations of Word Alignment Tool,” in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, ser. SETQA-NLP ’08, Columbus, OH, USA, June 2008, pp. 49–57.

[23] F. J. Och and H. Ney, “A Systematic Comparison of Various Statistical Alignment Models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[24] C. Dyer, V. Chahuneau, and N. A. Smith, “A Simple, Fast, and Effective Reparameterization of IBM Model 2,” in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Atlanta, GA, USA, June 2013, pp. 644–648.

[25] A. Stolcke, “SRILM – an Extensible Language Modeling Toolkit,” in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Denver, CO, USA, Sept. 2002, pp. 901–904.

[26] S. F. Chen and J. Goodman, “An Empirical Study of Smoothing Techniques for Language Modeling,” Computer Science Group, Harvard University, Cambridge, MA, USA, Tech. Rep. TR-10-98, Aug. 1998.

[27] K. Heafield, “KenLM: Faster and Smaller Language Model Queries,” in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, United Kingdom, July 2011, pp. 187–197.

[28] F. J. Och and H. Ney, “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation,” in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002, pp. 295–302.

[29] C. Cherry and G. Foster, “Batch Tuning Strategies for Statistical Machine Translation,” in *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, Montreal, Canada, June 2012, pp. 427–436.

[30] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a Method for Automatic Evaluation of Machine Translation,” in *Proc. of the 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Philadelphia, PA, USA, July 2002, pp. 311–318.

[31] M. Freitag, J. Wuebker, S. Peitz, H. Ney, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani, I. Slawik, J. Niehues, E. Cho, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, “Combined Spoken Language Translation,” in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, South Lake Tahoe, CA, USA, Dec. 2014.

[32] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, “Fast and robust neural network joint models for statistical machine translation,” in *52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA, June 2014.

[33] N. Durrani, A. Fraser, H. Schmid, H. Hoang, and P. Koehn, “Can markov models over minimal translation units help phrase-based smt?” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 399–405. [Online]. Available: http://www.aclweb.org/anthology/P13-2071

[34] N. Durrani, A. Fraser, and H. Schmid, “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, June 2013, pp. 1–11.

[35] D. Chiang, K. Knight, and W. Wang, “11,001 new features for statistical machine translation,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder, Colorado, June 2009, pp. 218–226.

[36] E. Matusov, A. Mauser, and H. Ney, “Automatic sentence segmentation and punctuation prediction for spoken language translation,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, November 2006.

[37] M. Auli, M. Galley, C. Quirk, and G. Zweig, “Joint language and translation modeling with recurrent neural networks.” in *EMNLP*, 2013, pp. 1044–1054.

[38] J. Gao, X. He, W.-t. Yih, and L. Deng, "Learning continuous phrase representations for translation modeling," in *Proc. ACL*, 2014.

[39] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, "Decoding with Large-Scale Neural Language Models Improves Translation," in *EMNLP*, 2013, pp. 1387–1392.

[40] P. Baltescu, P. Blunsom, and H. Hoang, "OxLM: A Neural Language Modelling Framework for Machine Translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 102, no. 1, pp. 81–92, 2014.

[41] N. Durrani, H. Sajjad, H. Hoang, and P. Koehn, "Integrating an Unsupervised Transliteration Model into Statistical Machine Translation," in *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2014)*. Gothenburg, Sweden: Association for Computational Linguistics, April 2014.

[42] H. Sajjad, A. Fraser, and H. Schmid, "A Statistical Model for Unsupervised and Semi-supervised Transliteration Mining," in *ACL12*, Jeju, Korea, 2012.

[43] N. Habash and F. Sadat, "Arabic Preprocessing Schemes for Statistical Machine Translation," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City, USA, June 2006, pp. 49–52.

[44] A. Axelrod, X. He, and J. Gao, "Domain Adaptation via Pseudo In-Domain Data Selection," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, July 2011, pp. 355–362.

[45] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes," in *Ninth Conference the European Chapter of the Association for Computational Linguistics (EACL)*, June 1999, pp. 71–76.

[46] M. Collins, P. Koehn, and I. Kucerova, "Clause Restructuring for Statistical Machine Translation," in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan, June 2005, pp. 531–540.

[47] P. Koehn and K. Knight, "Empirical methods for compound splitting," in *Proceedings of Meeting of the European Chapter of the Association of Computational Linguistics (EACL)*, 2003.

[48] H. Schmid, "Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors," in *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, Geneva, Switzerland, Aug. 2004.

[49] A. Bisazza, N. Ruiz, and M. Federico, "Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, USA, Dec. 2011, pp. 136–143.

[50] M. Huck, H. Hoang, and P. Koehn, "Augmenting String-to-Tree and Tree-to-String Translation with Non-Syntactic Phrases," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 486–498.

[51] ——, "Preference Grammars and Soft Syntactic Constraints for GHKM Syntax-based Statistical Machine Translation," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, Oct. 2014, pp. 148–156.

[52] M. Freitag, M. Huck, and H. Ney, "Jane: Open Source Machine Translation System Combination," in *Proc. of the Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL)*, Gothenburg, Sweden, Apr. 2014, pp. 29–32.