The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015

Barry Haddow[1], Matthias Huck[1], Alexandra Birch[1],
Nikolay Bogoychev[1], Philipp Koehn[1,2]
[1]School of Informatics, University of Edinburgh
[2]Center for Speech and Language Processing, The Johns Hopkins University
a.birch@ed.ac.uk {nbogoych,bhaddow,mhuck}@inf.ed.ac.uk phi@jhu.edu

Abstract

This paper describes the submission of the University of Edinburgh and the Johns Hopkins University for the shared translation task of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation (WMT 2015). We set up phrase-based statistical machine translation systems for all ten language pairs of this year's evaluation campaign, which are English paired with Czech, Finnish, French, German, and Russian in both translation directions.

Novel research directions we investigated include: neural network language models and bilingual neural network language models, a comprehensive use of word classes, and sparse lexicalized reordering features.

1 Introduction

The Edinburgh/JHU phrase-based translation systems for our participation in the WMT 2015 shared translation task[1] are based on the open source Moses toolkit (Koehn et al., 2007). We built upon Edinburgh's strong baselines from WMT submissions in previous years (Durrani et al., 2014a) as well as our recent research within the framework of other evaluation campaigns and projects such as IWSLT[2] and EU-BRIDGE[3] (Birch et al., 2014; Freitag et al., 2014a; Freitag et al., 2014b).

We first discuss novel features that we integrated into our systems for the 2015 Edinburgh/JHU submission. Next we give a general system overview with details on our training pipeline and decoder configuration. We finally present empirical results for the individual language pairs and translation directions.

[1]http://www.statmt.org/wmt15/
[2]http://workshop2014.iwslt.org
[3]http://www.eu-bridge.eu

2 Novel Methods

2.1 Neural Network LM with NPLM

For some language pairs (notably French↔English and Finnish↔English) we experimented with feed-forward neural network language models using the NPLM toolkit (Vaswani et al., 2013). This toolkit enables such language models to be trained efficiently on large datasets, and provides a querying API which is fast enough to be used during decoding. NPLM is fully integrated into Moses, including appropriate wrapper scripts for training the language models within the Moses experiment management system.

2.2 Bilingual Neural Network LM

We also experimented with our re-implementation of the "joint" model by Devlin et al. (2014). Referred to as *bilingual LM* in Moses, this was previously employed in the Edinburgh IWSLT system submissions, although with limited success (Birch et al., 2014).

The idea of the bilingual LM is quite straightforward. We define a language model where each target token is conditioned on the previous $(n-1)$ target tokens (as in a standard n-gram language model) as well as its aligned source token, and a window of m tokens on either side of the aligned source token. At training time, the aligned source token is found from the automatic alignment, and at test time the alignment is supplied by the decoder. The bilingual LM is trained using a feed-forward neural network and we use the NPLM toolkit for this.

Prior to submission we tested bilingual LMs on the French↔English tasks and on English→Russian task. For French↔English, we had resource issues[4] in training such large

[4]These can now be addressed using the `-mmap` option to create a binarized version of the corpus which is then memory-mapped.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 126–133,
Lisboa, Portugal, 17-18 September 2015.

models so we randomly subsampled 10% of the data for training. Since we did not observe gains in translation quality, the bilingual LM was not integrated into our primary system submissions. In post-submission experiments, we tried training bilingual LM on a 10% domain-specific portion of the training data selected using modified Moore-Lewis (Moore and Lewis, 2010; Axelrod et al., 2011), but only observed a small improvement in translation performance.

2.3 Comprehensive Use of Word Classes

In Edinburgh's submission from the previous year, we used automatically generated word classes in additional language models and in additional operation sequence models (Durrani et al., 2014b). This year, we pushed the use of word classes into the remaining feature functions: the reordering model and the sparse word features.

We generated Och clusters (Och, 1999) — a variant of Brown clusters — using `mkcls`. We have to choose a hyper parameter: the number of clusters. Our experiments and also prior work (Stewart et al., 2014) suggest that instead of committing to a single value, it is beneficial to use multiple numbers and use them in multiple feature functions concurrently. We used 50, 200, 600, and 2000 clusters, hence having 4 additional interpolated language models, 4 additional operation sequence models, 4 additional lexicalized reordering models, and 4 additional sets of sparse features.

The feature functions for word classes were trained exactly the same way as the corresponding feature functions for words. For instance, this means that the word class language model required training of individual models on the subcorpora, and then interpolation.

We carried out a study to assess the contribution of the use of such word class feature functions. Table 1 summarizes the results. Use of word classes in each of the models yields small gains, except for the reordering model, where there is no observable difference. The biggest gains were observed in the language model. Note that the English–German baseline already included additional feature functions based on POS and morphological tags, and basically no additional gains were observed due to the class based feature functions.

2.4 Sparse Lexicalized Reordering

We implemented sparse lexicalized reordering features (Cherry, 2013) in Moses and evaluated them in English↔German setups. The experiments were conducted on top of the standard hierarchical lexicalized reordering model (Galley and Manning, 2008). We applied features based on Och clusters with 200 classes on both source and target side. Active feature groups are *between*, *phrase*, and *stack*.

In addition to optimizing the feature weights directly with k-best MIRA (Cherry and Foster, 2012), we also examined maximum expected BLEU training of the sparse lexicalized reordering features via stochastic gradient descent (Auli et al., 2014).

3 System Overview

3.1 Preprocessing

The training data was preprocessed using scripts from the Moses toolkit. We first normalized the data using the `normalize-punctuation.perl` script, then performed tokenization (using the `-a` option), and then truecasing. We did not perform any corpus filtering other than the standard Moses method, which removes sentence pairs with extreme length ratios.

3.2 Word Alignment

For word alignment we used either `fast_align` (Dyer et al., 2013) or MGIZA++ (Gao and Vogel, 2008), followed by the standard `grow-diag-final-and` symmetrization heuristic. An empirical comparison of `fast_align` and MGIZA++ on the Finnish-English and English-Russian language pairs using the constrained data sets did not reveal any significant difference.

3.3 Language Model

We used all available monolingual data to train 5-gram language models with modified Kneser-Ney smoothing (Chen and Goodman, 1998). Typically, language models for each monolingual corpus were first trained using either KenLM (Heafield et al., 2013) or the SRILM toolkit (Stolcke, 2002) and then linearly interpolated using weights tuned to minimize perplexity on the development set.

3.4 Baseline Features

We follow the standard approach to SMT of scoring translation hypotheses using a weighted linear combination of features. The core features of our

	de-en	en-de	cs-en	en-cs	ru-en	en-ru	avg Δ
Baseline (no clusters)	28.0	20.5	29.1	21.2	31.8	29.1	-
Comprehensive setup	28.5 (+.5)	20.5 (±.0)	29.7 (+.6)	21.8 (+.6)	32.3 (+.5)	29.7 (+.6)	+.5
w/o sparse features	28.2 (–.3)	20.4 (–.1)	29.6 (–.1)	21.7 (–.1)	32.2 (–.1)	30.0 (+.3)	–.2
w/o language model	28.3 (–.2)	20.5 (±.0)	29.5 (–.2)	21.4 (–.4)	31.5 (–.8)	29.2 (–.6)	–.4
w/o reordering model	28.5 (±.0)	20.5 (±.0)	-	21.8 (±.0)	32.3 (±.0)	29.8 (+.1)	±.0
w/o operation sequence model	28.3 (–.2)	20.3 (–.1)	29.7 (±.0)	21.7 (–.1)	32.0 (–.3)	29.5 (–.2)	–.2

Table 1: Use of additional feature functions based on Och clusters (see Section 2.3). The last four lines refer to ablation studies where one of the sets of clustered feature functions is removed from the comprehensive setup. Note that the word-based feature functions are used in all cases. BLEU scores on `newstest2014` are reported.

model are a 5-gram LM score, phrase translation and lexical translation scores, word and phrase penalties, and a linear distortion score. The phrase translation probabilities are smoothed with Good-Turing smoothing (Foster et al., 2006). We used the hierarchical lexicalized reordering model (Galley and Manning, 2008) with 4 possible orientations (monotone, swap, discontinuous left and discontinuous right) in both left-to-right and right-to-left direction. We also used the operation sequence model (OSM) (Durrani et al., 2013) with 4 count based supportive features. We further employed domain indicator features (marking which training corpus each phrase pair was found in), binary phrase count indicator features, sparse phrase length features, and sparse source word deletion, target word insertion, and word translation features (limited to the top K words in each language, typically with $K = 50$).

3.5 Tuning

Since our feature set (generally around 500 to 1000 features) was too large for MERT, we used k-best batch MIRA for tuning (Cherry and Foster, 2012). To speed up tuning we applied threshold pruning to the phrase table, based on the direct translation model probability.

3.6 Decoding

In decoding we applied cube pruning (Huang and Chiang, 2007) with a stack size of 5000 (reduced to 1000 for tuning), Minimum Bayes Risk decoding (Kumar and Byrne, 2004), a maximum phrase length of 5, a distortion limit of 6, 100-best translation options and the no-reordering-over-punctuation heuristic (Koehn and Haddow, 2009).

4 Experimental Results

In this section we describe peculiarities of individual systems and present experimental results.

4.1 French↔English

Our submitted systems for the French-English language pair are quite similar for the two translation directions. We used all the constrained parallel data to build a phrase-based translation model and the language model was build from the target side of this data, the monolingual news data and the LDC GigaWord corpora. During system development we used the `newsdiscussdev2015` for tuning and development testing, using 2-fold cross validation. For tuning the submitted system, and the post-submission experiments, we tuned on the whole of `newsdiscussdev2015`, and report cased BLEU on `newsdiscusstest2015`.

Prior to submission we experimented with bilingual LM and an NPLM-based neural network language model (Sections 2.2 and 2.1) but did not obtain positive results. These were trained on a randomly selected 10% portion of the parallel training data. We also experimented with class-based language models (using Och clusters from `mkcls`), including the 50 class language model in the English→French submission but not in the French→English one, since it helped in our development setup in the former but not the latter.

In the post-submission experiments (Table 2), we show the comparison of the baseline system (as described in Section 3) with systems enhanced with bilingual LM, NPLM and class-based language models. For the class-based language models, we tested with 50 Och clusters, 200 Och clusters, and with both class-based LMs. For the bilingual LM, we created both "combined" (a 5-gram on the target and a 9-gram on the source) and "source" (1-gram on the target and 15-gram on

System	fr-en	en-fr
Baseline	33.0	33.5
Submitted	32.7	33.6
50 classes	32.8	33.8
200 classes	32.9	33.9
50+200 classes	32.9	33.7
BiLM combined	32.9	33.6
BiLM source & combined	33.2	33.5
NPLM	33.0	34.2

Table 2: Comparison of baseline with post-submission experiments on class-based language models, bilingual LM and NPLM. Note that for French→English the submitted system was the same as the baseline (retuned) whilst for English→French it was the same as the third line (retrained).

System	fi-en	en-fi
Baseline	19.6	13.4
Submitted	19.7	n/a
Without OPUS	17.0	11.5
50 classes	19.4	13.2
200 classes	19.8	13.3
50+200 classes	19.7	13.3
BiLM combined	19.1	13.5
BiLM source & combined	19.1	13.4
NPLM	20.0	13.8

Table 3: Comparison of baseline with post-submission experiments on class-based language models, bilingual LM and NPLM. Note that the submitted system for Finnish→English was the same as the baseline (but retuned).

source) models. The bilingual LMs are trained on 10% of the available parallel data, selected using modified Moore-Lewis data selection (Moore and Lewis, 2010; Axelrod et al., 2011). The NPLM is a 5-gram model trained on all available language model data.

We observe from Table 2 that the bilingual LM has a minimal effect on BLEU, only showing an increase for one language pair, one configuration, and the margin of improvement is probably within the margin of tuning variation. We do not have a good explanation for the lack of success with bilingual LM, in contrast to (Devlin et al., 2014), however we note that all reports of improvements with this type of model are for distantly related language pairs. We also did not observe any improvement with the class-based language models for French→English, although we did observe small gains from English→French. Building an NPLM model for all data gives a reasonable improvement (+0.7) for the French target, but not the English. In fact French→English was the only language pair where NPLM did not improve BLEU after building the LM on all data. It is possible that the limited morphology of English means that the improved generalisation of the NPLM is not as helpful, and also that the conventional n-gram LM is already strong for this language pair.

4.2 Finnish↔English

For the Finnish-English language pair we built systems using only the constrained data, and systems using all the OPUS (Tiedemann, 2012) parallel data. Our baselines include this extra data, but we also show results just using the constrained parallel data. We did not employ the morphological splitting as in Edinburgh's syntax-based system (Williams et al., 2015) and consequently the English→Finnish systems performed poorly in development and we did not submit a phrase-based system for this pair.

Our development setup was similar to French↔English; we used the `newsdev2015` for tuning and test during system development (in 2-fold cross-validation) then for the submission and subsequent experiments we used the whole of `newsdev2015` for tuning. Also in common with our work on French↔English, we performed several post-submission experiments to examine the effect of class-based language models, bilingual LM and NPLM. We show the results in Table 3. For training bilingual LM and NPLM models we encountered some numerical issues, probably due to the large vocabulary size in Finnish. These were partially addressed by employing *dropout* to prevent overfitting (Srivastava et al., 2014), enabling us to train the models for at least 2 epochs.

We note that, as with French↔English, our application of bilingual LM did not result in significant improvement. Finnish and English are quite distantly related, but we can speculate that using words as a representation for Finnish is not appropriate. The NPLM, however, offers modest (+0.4) improvements over the baseline in both directions.

4.3 Czech↔English

The development of the Czech↔English systems followed the ideas in Section 2.3, i.e., with a focus on word classes (50, 200, 600 classes) for all component models. We combined the test sets from 2008 to 2012 for tuning. No neural language model or bilingual language model was used.

4.4 Russian↔English

To Russian. For the English→Russian system, we used all the parallel data specified in the task. The Wiki Headlines data was appended onto the combined parallel corpus. For the monolingual corpora, we used all the constrained track corpora except for Newscrawl 2008-2010 which were overlooked as they were much smaller than other resources. We trained word classes with three different settings (50, 200, and 600 clusters) on both source and target languages. On applying clusters, we trained 6-gram language models on the target side. We used all four factors (words and clusters) in both source and target languages for the the translation model and the OSM, but we used only the word factor for the alignment and the reordering models. We performed transliteration (Durrani et al., 2014c) after decoding for all three experimental conditions. We used `newstest2012` for LM interpolation and batch MIRA model tuning. In Table 4, the only difference between the baseline system and the official submission is that the baseline has no cluster factors. The final model (BiLM source & combined & NPLM) is the same as the submitted system, apart from the fact that we applied two bilingual neural network models: one over the source and one over the source and target, and an NPLM language model over the target. This did not improve over the factored model and so was not submitted for the evaluation.

From Russian. The Russian→English system used the same settings as the Czech system, except for the addition of a factor over 2000 word classes and a smaller tuning set (just `newstest2012`).

4.5 German↔English

Our German-English training corpus comprises all permissible parallel data of the constrained track for this language pair. A concatenation of `newssyscomb2009` and `newstest2008-2012` served as tuning set.

System	en-ru
Baseline	25.0
Submitted	25.2
BiLM source & combined & NPLM	25.1

Table 4: Experimental results (cased BLEU) for English→Russian averaged over `newstest2013` and `newstest2014`.

From German. For translation from German, we applied syntactic pre-reordering (Collins et al., 2005) and compound splitting (Koehn and Knight, 2003) in a preprocessing step on the source side. A rich set of translation factors was exploited in addition to word surface forms: Och clusters (50 classes), morphological tags, part-of-speech tags, and word stems on the German side (Schmid, 2000), as well as Och clusters (50 classes), part-of-speech tags (Ratnaparkhi, 1996), and word stems (Porter, 1980) on the English side. The factors were utilized in the translation model and in OSMs. The lexicalized reordering model was trained on stems. Individual 7-gram Och cluster LMs were trained with KenLM's `--discount_fallback --prune '0 0 1'` parameters,[5] then interpolated with the SRILM toolkit and added to the log-linear model as a second LM feature. Our 5-gram word LM was trained on all English data at once, also with pruning of singleton n-grams of order 3 and higher. We included the English LDC Gigaword Fifth Edition. Sparse lexical features (source word deletion, target word insertion, word translation) were limited to the top $K = 200$ words for German→English.

To German. Translation factors for the English→German translation direction are word surface forms, Och clusters (50 classes), morphological tags, and part-of-speech tags. Morphological tags were employed on the target side only, all other factors on both source and target side. The lexicalized reordering model was trained on word surface forms. We added an interpolated 7-gram Och cluster LM and a 7-gram LM over morphological tags. LMs were trained in a similar way as the ones for translation from German. Sparse phrase length features and sparse lexical features were not used for English→German.

[5] `http://www.statmt.org/mtm14/uploads/Projects/KenLMFunWithLanguageModel_MTM2014p9.pdf`

System	de-en		en-de	
	2013	2014	2013	2014
Baseline	27.3	28.6	20.6	20.9
+ sparse LR (MIRA)	**27.2**	**28.8**	20.7	20.8
+ sparse LR (SGD)	27.2	28.5	**20.8**	**21.1**

Table 5: Experimental results for German→English and English→German. We report cased BLEU scores on the newstest2013 and newstest2014 sets. Primary submission results are highlighted in bold.

Sparse lexicalized reordering. We investigated sparse lexicalized reordering features (Section 2.4) on the German-English language pair in both translation directions. Two methods for learning the weights of the sparse lexicalized reordering feature set have been compared: (1.) direct tuning in MIRA along with all other features in the model combination (*sparse LR (MIRA)*), and (2.) separate optimization with stochastic gradient descent (SGD) with a maximum expected BLEU objective (*sparse LR (SGD)*). For the latter variant, we used the MT tuning set for training (13 573 sentence pairs) and otherwise followed the approach outlined by Auli et al. (2014). We tuned the baseline feature weights with MIRA before SGD training and ran two final MIRA iterations after it. SGD training was stopped after 80 epochs.

Empirical results for the German-English language pair are presented in Table 5. We observe minor gains of up to +0.2 points BLEU. The results are not consistent in the two translation directions: The MIRA-trained variant seems to perform better when translating from German, the SGD-trained variant when translating to German. However, in both cases the baseline score is almost identical to the best results with sparse lexicalized reordering features.

In future work we plan to adopt hypergraph MIRA, as well as larger training sets for maximum expected BLEU training. We also consider scaling the method to word surface forms in addition to Och clusters, and trying RPROP instead of SGD.

5 Conclusion

The Edinburgh/JHU team built phrase-based translation systems using the open source Moses toolkit for all language pairs of the WMT 2015 shared translation task. Our submitted system outputs ranked first according to cased BLEU on the newstest2015 evaluation set on six out of ten language pairs:[6] Czech→English, German→English, Finnish→English, Russian→English, English→French, and English→Russian.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), 645487 (MMT), 644333 (TraMOOC) and 644402 (HimL).

References

Michael Auli, Michel Galley, and Jianfeng Gao. 2014. Large-scale Expected BLEU Training of Phrase-based Reordering Models. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1250–1260, Doha, Qatar, October.

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.

Alexandra Birch, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, and Philipp Koehn. 2014. Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, pages 49–56, Lake Tahoe, CA, USA, December.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

Colin Cherry. 2013. Improved Reordering for Phrase-Based Translation using Sparse Features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 22–31, Atlanta, GA, USA, June.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual*

[6] http://matrix.statmt.org/?mode=all&test_set[id]=21

Meeting of the Association for Computational Linguistics (ACL'05), pages 531–540, Ann Arbor, MI, USA, June.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, MD, USA, June.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013. 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)*, pages 399–405, Sofia, Bulgaria, August.

Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. 2014a. Edinburgh's Phrase-based Machine Translation Systems for WMT-14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 97–104, Baltimore, MD, USA, June.

Nadir Durrani, Philipp Koehn, Helmut Schmid, and Alexander Fraser. 2014b. Investigating the Usefulness of Generalized Word Representations in SMT. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 421–432, Dublin, Ireland, August.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014c. Integrating an unsupervised transliteration model into statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 148–153, Gothenburg, Sweden, April. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, GA, USA, June.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable Smoothing for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 53–61, Sydney, Australia, July.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014a. EU-BRIDGE MT: Combined Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 105–113, Baltimore, MD, USA, June.

Markus Freitag, Joern Wuebker, Stephan Peitz, Hermann Ney, Matthias Huck, Alexandra Birch, Nadir Durrani, Philipp Koehn, Mohammed Mediani, Isabel Slawik, Jan Niehues, Eunah Cho, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2014b. Combined Spoken Language Translation. In *International Workshop on Spoken Language Translation*, pages 57–64, Lake Tahoe, CA, USA, December.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu, HI, USA.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Liang Huang and David Chiang. 2007. Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic.

Philipp Koehn and Barry Haddow. 2009. Edinburgh's Submission to all Tracks of the WMT 2009 Shared Task with Reordering and Speed Improvements to Moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 160–164, Athens, Greece.

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 187–194, Budapest, Hungary, April.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *HLT-NAACL 2004: Main Proceedings*, pages 169–176, Boston, MA, USA.

Robert C. Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden.

Franz Josef Och. 1999. 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)*, pages 71–76.

Martin Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, May.

Helmut Schmid. 2000. LoPar: Design and Implementation. Bericht des Sonderforschungsbereiches "Sprachtheoretische Grundlagen für die Computerlinguistik" 149, Institute for Computational Linguistics, University of Stuttgart.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Darlene Stewart, Roland Kuhn, Eric Joanis, and George Foster. 2014. Coarse split and lump bilingual language models for richer source information in SMT. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1, pages 28–41.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, volume 3, Denver, CO, USA, September.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, pages 2214–2218, Istanbul, Turkey, May.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392, Seattle, WA, USA.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the EMNLP 2015 Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September.