The Edinburgh/LMU Hierarchical Machine Translation System for WMT 2016

Matthias Huck[1], Alexander Fraser[1], Barry Haddow[2]
[1]Center for Information and Language Processing, LMU Munich
[2]School of Informatics, University of Edinburgh
mhuck@cis.lmu.de fraser@cis.lmu.de bhaddow@inf.ed.ac.uk

Abstract

This paper describes the hierarchical phrase-based machine translation system built jointly by the University of Edinburgh and the University of Munich (LMU) for the shared translation task at the ACL 2016 First Conference on Machine Translation (WMT16). The WMT16 Edinburgh/LMU system was trained for translation of news domain texts from English into Romanian. We participated in the shared task for machine translation of news under "constrained" conditions, i.e. using the provided training data only.

1 Introduction

While translation between English and many other European languages (such as Czech and German) has a long tradition in the shared tasks at the series of WMT workshops preceding the ACL 2016 First Conference on Machine Translation, English–Romanian has only been introduced this year as a new language pair.[1] The English–Romanian language pair has received less attention by the machine translation scientific community to date. The availability of a novel standardized evaluation scenario for English–Romanian in the framework of WMT facilitates research on that specific language pair.

In this work, we utilize the corpora that have been provided by the shared task organizers to engineer a competitive system for statistical machine translation (SMT) from English into Romanian. We specifically focus on studying machine translation into Romanian (rather than the inverse translation direction: from Romanian into English), thus aiming at making documents originally written in English available to a large community of speakers in their native language, Romanian. Applications are for instance in the health care sector, where, as part of the *Health in my Language* project (*HimL*), several project partners intend to make public health information available in a wider variety of languages.[2] The WMT task provides an interesting test bed for English→Romanian machine translation, though adaptation towards the specific domain (consumer health for *HimL*, rather than news) is also an important aspect that has to be considered in practice (Huck et al., 2015).

We investigate the effectiveness of *hierarchical phrase-based translation* (Chiang, 2005) for English→Romanian, a statistical machine translation paradigm that is closely related to phrase-based translation, but allows for phrases with gaps. Conceptionally, the translation model is formalized as a synchronous context-free grammar. We integrate several non-standard enhancements into our hierarchical phrase-based system and empirically evaluate their impact on translation quality.

Our system is furthermore one component in a combination of systems by members of the *HimL* project and another EU-funded project, *QT21*.[3] Measured in BLEU (Papineni et al., 2002), the *QT21/HimL* submission yields top translation quality amongst the shared task submissions.[4] The *QT21/HimL* submission highlights the continued success of system combinations based on the Jane machine translation toolkit (Freitag et al., 2014a) in open evaluation campaigns (Freitag et al., 2013; Freitag et al., 2014b; Freitag et al., 2014c). A description of the *QT21/HimL* combined submission is given by Peter et al. (2016).

[1] http://www.statmt.org/wmt16/translation-task.html

[2] http://www.himl.eu

[3] http://www.qt21.eu

[4] http://matrix.statmt.org/matrix/systems_list/1843

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 311–318,
Berlin, Germany, August 11-12, 2016.

We proceed by presenting the particularities of our hierarchical phrase-based system, with a focus of interest on exploring non-standard enhancements and non-default configuration settings such as:

- Individual language models as features, rather than a single linearly interpolated language model; and another background language model estimated over concatenated corpora.
- Large CommonCrawl language model training data.
- Unpruned language models.
- More hierarchical rules than in default systems, by means of imposing less strict extraction constraints.
- A phrase orientation model for hierarchical translation (Huck et al., 2013).
- Lightly-supervised training (Schwenk, 2008; Schwenk and Senellart, 2009; Huck et al., 2011).
- Larger development data for tuning.

All our experiments are run with the open source `Moses` implementation (Hoang et al., 2009) of the hierarchical phrase-based translation paradigm.

2 System Overview

2.1 Hierarchical Phrase-Based Translation

In hierarchical phrase-based translation, a probabilistic synchronous context-free grammar is induced from bilingual training corpora. In addition to continuous *lexical* phrases as in standard phrase-based translation, *hierarchical* phrases with (usually) up to two non-terminals are extracted from the word-aligned parallel training data.

The non-terminal set of a standard hierarchical grammar comprises two symbols which are shared by source and target: the initial symbol S and one generic non-terminal symbol X. The initial symbol S is the start symbol of the grammar. The generic non-terminal X is used as a placeholder for the gaps within the right-hand side of hierarchical translation rules as well as on all left-hand sides of the translation rules that are extracted from the parallel training corpus.

Extracted rules of a standard hierarchical grammar are of the form $X \rightarrow \langle \alpha, \beta, {}^{\sim} \rangle$ where $\langle \alpha, \beta \rangle$ is a bilingual phrase pair that may contain X, i.e. $\alpha \in (\{X\} \cup V_F)^+$ and $\beta \in (\{X\} \cup V_E)^+$, where V_F and V_E are the source and target vocabulary, respectively. The non-terminals on the source side and on the target side of hierarchical rules are linked in a one-to-one correspondence. The ${}^{\sim}$ relation defines this one-to-one correspondence.

In addition to the extracted rules, a non-lexicalized *glue rule*

$$S \rightarrow \left\langle S^{\sim 0} X^{\sim 1}, S^{\sim 0} X^{\sim 1} \right\rangle \tag{1}$$

is incorporated into the hierarchical grammar that the system can use for serial concatenation of phrases as in monotonic phrase-based translation.

In the `Moses` implementation, the decoder internally adds a *sentence start* terminal symbol <s> and a *sentence end* terminal symbol </s> to the input before and after each sentence, respectively. Therefore, two more special rules

$$\begin{aligned} S &\rightarrow \left\langle \text{<s>}, \text{<s>} \right\rangle \\ S &\rightarrow \left\langle S^{\sim 0} \text{</s>}, S^{\sim 0} \text{</s>} \right\rangle \end{aligned} \tag{2}$$

are included which allow the decoder to finalize its translations.

Hierarchical search is conducted with a customized version of the CYK+ parsing algorithm (Chappelier and Rajman, 1998) and cube pruning (Chiang, 2007). A hypergraph which represents the whole parsing space is built employing CYK+. Cube pruning operates in bottom-up topological order on this hypergraph and expands at most k derivations at each hypernode.

2.2 Data and Preprocessing

Our system is trained using only permissible Romanian monolingual and English–Romanian parallel corpora provided by the organizers of the WMT16 shared task for machine translation of news: Europarl (Koehn, 2005), SETimes2 (Tyers and Alperen, 2010), News Crawl articles from 2015 (denoted as news2015 hereafter), and CommonCrawl (Buck et al., 2014).

The target side of the data is preprocessed with `tokro`, LIMSI's tokenizer for Romanian (Allauzen et al., 2016).[5] The English source side is tokenized using the `tokenizer.perl` script from the `Moses` toolkit. Romanian and English sentences are both frequent-cased (with `Moses`' `truecase.perl`).

[5] `https://perso.limsi.fr/aufrant/software/tokro`

We split the development set newsdev2016 into two halves (newsdev2016_1 with the first 1000 sentences and newsdev2016_2 with the last 999 sentences). During the system building process, we measure progress by evaluating on newsdev2016_2 as our internal unseen test set, while only newsdev2016_1 is utilized for tuning.

2.3 Training and Tuning

We create word alignments by aligning the bilingual data in both directions with `MGIZA++` (Gao and Vogel, 2008). We use a sequence of IBM word alignment models (Brown et al., 1993) with five iterations of EM training (Dempster et al., 1977) of Model 1, three iterations of Model 3, and three iterations of Model 4. After EM, we obtain a symmetrized alignment by applying the `grow-diag-final-and` heuristic (Och and Ney, 2003; Koehn et al., 2003) to the two trained alignments. We extract synchronous context-free grammar rules that are consistent with the symmetrized word alignment from the parallel training data.

We train 5-gram language models (LMs) with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). `KenLM` (Heafield, 2011) is employed for LM training and scoring, and `SRILM` (Stolcke, 2002) for linear LM interpolation.

Our translation model incorporates a number of different features in a log-linear combination (Och and Ney, 2002). We tune the feature weights with batch *k*-best MIRA (Cherry and Foster, 2012) to maximize BLEU (Papineni et al., 2002) on a development set. We run MIRA for 25 iterations on 200-best lists.

2.4 Baseline Setup

The features of our plain hierarchical phrase-based baseline are:

- Rule translation log-probabilities in both target-to-source and source-to-target direction, smoothed with Good-Turing discounting (Foster et al., 2006).
- Lexical translation log-probabilities in both target-to-source and source-to-target direction.
- Seven binary features indicating absolute occurrence count classes of translation rules (with count classes 1, 2, 3, 4, 5-6, 7-10, >10).
- An indicator feature that fires on applications of the glue rule.
- Word penalty.
- Rule penalty.
- A 5-gram language model.

We discard rules with non-terminals on their right-hand side if they are singletons in the training data. The baseline language model is a linear interpolation of three 5-gram LMs trained over the Romanian news2015, Europarl, and SETimes2 training data, respectively, with pruning of singleton *n*-grams of order three and higher.[6] We run the `Moses` chart-based decoder with cube pruning, configured at a maximum chart span of 25 and otherwise default settings.

2.5 Enhancements

We now describe modifications that we apply on top of the baseline. The results of the empirical evaluation will be given in Section 3.

Linear LM interpolation vs. individual LMs as features in the log-linear combination. Rather than employing a linearly interpolated LM, we integrate the individual LMs trained over the separate corpora (news2015, Europarl, SETimes2) directly into the log-linear feature combination of the system and let MIRA optimize their weights along with all other features in tuning.

Background LM. We add one more language model, which we denote as *background LM*. The background LM is estimated from a concatenation of the Romanian news2015, Europarl, and SETimes2 training data. The background LM does not replace the individual LMs in the log-linear combination, but acts as another feature with an associated weight.

CommonCrawl LM training data. A large Romanian CommonCrawl corpus has been released for the constrained track of the WMT16 shared task for machine translation of news. In our system, we utilize this corpus by adding it to the training data of the background LM. We append it to the concatenation of news2015, Europarl, and SETimes2 data and estimate a bigger background LM.

[6] Pruned individual LMs are trained with KenLM's `--prune '0 0 1'` parameters. Weights for linear LM interpolation are optimized on newsdev2016_1.

Pruned vs. unpruned LMs. We compare pruned and unpruned language models. In the pruned versions of the models, singleton *n*-grams of order three and higher are discarded, whereas all *n*-grams are kept in the unpruned versions.

More hierarchical rules. The baseline synchronous context-free grammar rules in the phrase table are extracted from the parallel training data with `Moses`' default settings: a maximum of five symbols on the source side, a maximum span of ten words, and no right-hand side non-terminal at gaps that cover only a single word on the source side. We allow for extraction of more hierarchical rules by applying less strict rule extraction constraints: a maximum of ten symbols on the source side, a maximum span of twenty words, and no lower limit to the amount of words covered by non-terminals at extraction time.

Phrase orientation model. We implemented a feature in `Moses` that resembles the phrase orientation model for hierarchical machine translation as described by Huck et al. (2013). The Huck et al. (2013) implementation had been released as part of the `Jane` toolkit (Vilar et al., 2010; Vilar et al., 2012; Huck et al., 2012). Our new `Moses` implementation technically operates in almost the same manner, except for minor implementation differences. Similarly to the type of lexicalized reordering models that are in common use in phrase-based systems (Galley and Manning, 2008), our model estimates the probabilities of orientation classes for each phrase (or: rule) from the training data. We use three orientation classes: *monotone*, *swap*, and *discontinuous*.[7]

Lightly-supervised training. We automatically translated parts (1.2 M sentences) of the monolingual Romanian news2015 corpus to English with a Romanian→English phrase-based statistical machine translation system (Williams et al., 2016). The resulting synthetic parallel corpus of the original Romanian news texts paired with machine-translated English counterparts is utilized for lightly-supervised training (Schwenk, 2008) of our English→Romanian hierarchical system.

We follow the approach outlined by Huck et al. (2011) to augment the system with the synthetic parallel data. A foreground phrase table extracted from the human-generated parallel data is filled up with entries from a background phrase table extracted from the synthetic parallel data. An entry from the background table is only added if the foreground table does not already contain a similar entry (Bisazza et al., 2011). A binary feature distinguishes background phrases from foreground phrases. For the background phrase table, we extract only lexical phrases (i.e., phrases without non-terminals on their right-hand side) from the synthetic parallel data, no hierarchical phrases. The phrase length for entries of the background table is restricted to a maximum number of five terminal symbols on the source side. Lexical scores over the phrases extracted from synthetic data are calculated with a lexicon model learned from the human-generated parallel data, as proposed by Huck and Ney (2012).

Larger development data. Since no dedicated unseen test set was available during system building, newsdev2016 was split into its first half (newsdev2016_1) and its second half (newsdev2016_2) so that we could tune on the first half and keep the second half untouched for evaluating progress in translation quality with the various enhancements. For the final system (our primary submission), we took the best configuration built in this manner and tuned it on both halves, i.e. all of newsdev2016. 1000 sentences (as in newsdev2016_1) are a relatively small size for a development set, and we suspected that the optimized feature weights could become more reliable with twice the amount of development data.[8] Good results when tuning on newsdev2016_1 and testing on newsdev2016_2 made us feel confident about keeping the overall system configuration fixed and re-tuning the feature weights on all of newsdev2016. We calculated the BLEU scores on newsdev2016_1 and newsdev2016_2 (both being part of the development set now) as a sanity check and then submitted a hypothesis translation for the evaluation set, newstest2016, without further internal validation on a test set.

[7] Using `Moses`' Experiment Management System (EMS) (Koehn, 2010), the phrase orientation model for hierarchical machine translation can be activated by simply adding a line `phrase-orientation = true` to the `[TRAINING]` section of the EMS configuration file.

[8] Whenever available, we typically attempt to use large development sets (in the order of a few thousand sentences), e.g. for Edinburgh's phrase-based systems for the German–English language pair (Haddow et al., 2015).

en→ro	newsdev2016_1	newsdev2016_2	newstest2016
baseline with interpolated LM over news2015, Europarl, SETimes2	22.1	26.6	23.0
+ three individual LMs (replacing the interpolated LM)	21.6	26.6	22.9
+ background LM over concatenation of news2015, Europarl, SETimes2	22.2	27.1	23.3
+ CommonCrawl LM training data in background LM	23.1	28.3	24.4
+ all LMs unpruned	23.4	28.6	24.4
+ more hierarchical rules	23.1	29.0	24.7
+ phrase orientation model	24.4	29.5	25.5
+ lightly-supervised training (*contrastive submission system*)	24.8	30.2	25.5
+ tuning on full newsdev2016 (*primary submission system*)	24.5	30.9	25.9

Table 1: Incremental improvements over a plain hierarchical phrase-based baseline for English→Romanian (case-sensitive BLEU scores). Feature weights are tuned on newsdev2016_1 in all experiments except the one in the bottom line, where both newsdev2016_1 and newsdev2016_2 are employed for tuning.

3 Experiments

Table 1 presents the results achieved with the plain hierarchical phrase-based baseline, and the gains when incrementally applying modifications as described in Section 2.5. The decoder output is postprocessed with the `detruecase.perl` script from the `Moses` toolkit for recasing and `tokro` with its `-r` command line switch for detokenization. We evaluate case-sensitive with `mteval-v13a.pl -c`.

3.1 Discussion

Replacing the baseline's linearly interpolated LM with three individual LMs as features in the log-linear combination deteriorates the BLEU score on the development set by half a point, but has barely any impact on translation quality on the test sets ($\pm$0.0 BLEU on newsdev2016_2, $-$0.1 BLEU on newstest2016). By also adding a background LM over the concatenated news2015, Europarl, and SETimes2 corpora, we attain a similar BLEU score on the development set as with the baseline's linearly interpolated LM, but a gain of +0.3 to +0.5 BLEU on the test sets, compared to the baseline.

Utilizing a larger amount of target-side monolingual resources by appending the CommonCrawl corpus to the background LM's training data is very beneficial and increases the BLEU scores by around one point. Not pruning the LMs, i.e. not discarding singleton *n*-grams of order three and higher, has a positive effect on newsdev2016_1 and newsdev2016_2 (+0.3 BLEU), but makes no difference on newstest2016. If we allow for extraction of more hierarchical rules, we slightly harm the result on the development set again, but the model seems to generalize better, with +0.4 BLEU on newsdev2016_2 and +0.3 BLEU on newstest2016.

The phrase orientation model performs particularly well on newstest2016, with a gain of another +0.8 BLEU. Lightly-supervised training, on the other hand, does not boost translation quality on newstest2016 at all, though we see a decent improvement on newsdev2016_2, our internal test set. (+0.7 BLEU).

In our very last experiment, when we tune on the concatenation of newsdev2016_1 and newsdev2016_2, we find that employing the larger development data is of benefit to the system (+0.4 BLEU on newstest2016).

Overall, the two individual system enhancements that give us the largest improvements on newstest2016 are the large Romanian CommonCrawl corpus (+1.1 BLEU) and the phrase orientation model (+0.8 BLEU).

4 Summary

We built a hierarchical phrase-based system for translation of news texts from English into Romanian. By enhancing the system with non-standard components, we have been able to achieve an overall improvement over a plain hierarchical baseline of +2.9 BLEU points on the newstest2016 set.

Our `Moses` reimplementation of the phrase orientation model for hierarchical machine translation (Huck et al., 2013) has been released as part of `Moses` on GitHub.[9]

[9] `https://github.com/moses-smt/mosesdecoder`

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 644402 (*HimL*).

We thank Franck Burlot and Lauriane Aufrant from LIMSI-CNRS in Orsay, France, for providing preprocessed corpora and sharing their Romanian tokenizer in the framework of a QT21/HimL cross-project collaboration.

References

Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, and François Yvon. 2016. LIMSI@WMT'16: Machine Translation of News. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA, USA, December.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram Counts and Language Models from the Common Crawl. In *Proc. of the Language Resources and Evaluation Conference*, Reykjavík, Iceland, May.

Jean-Cédric Chappelier and Martin Rajman. 1998. A Generalized CYK Algorithm for Parsing Stochastic CFG. In *Proc. of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137, Paris, France, April.

Stanley F. Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, USA, August.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 427–436, Montréal, Canada, June.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, MI, USA, June.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–22.

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

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Nadir Durrani, Matthias Huck, Philipp Koehn, Thanh-Le Ha, Jan Niehues, Mohammed Mediani, Teresa Herrmann, Alex Waibel, Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 128–135, Heidelberg, Germany, December.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014a. Jane: Open Source Machine Translation System Combination. In *Proc. of the Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 29–32, Gothenburg, Sweden, April.

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. 2014b. EU-BRIDGE MT: Combined Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, 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. 2014c. Combined Spoken Language Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 57–64, Lake Tahoe, CA, USA, December.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 847–855, Honolulu, HI, USA, October.

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, Columbus, OH, USA, June.

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 126–133, Lisbon, Portugal, September.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 187–197, Edinburgh, Scotland, UK, July.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 152–159, Tokyo, Japan, December.

Matthias Huck and Hermann Ney. 2012. Pivot Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, October.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proc. of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland, UK, July.

Matthias Huck, Jan-Thorsten Peter, Markus Freitag, Stephan Peitz, and Hermann Ney. 2012. Hierarchical Phrase-Based Translation with Jane 2. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 98:37–50, October.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A Phrase Orientation Model for Hierarchical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 452–463, Sofia, Bulgaria, August.

Matthias Huck, Alexandra Birch, and Barry Haddow. 2015. Mixed-Domain vs. Multi-Domain Statistical Machine Translation. In *Proc. of MT Summit XV, vol.1: MT Researchers' Track*, pages 240–255, Miami, FL, USA, October.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA, May.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 127–133, Edmonton, Canada, May/June.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proc. of the MT Summit X*, Phuket, Thailand, September.

Philipp Koehn. 2010. An Experimental Management System. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 94:87–96, September.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, USA, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, July.

Jan-Thorsten Peter, Tamer Alkhouli, Hermann Ney, Matthias Huck, Fabienne Braune, Alexander Fraser, Aleš Tamchyna, Ondřej Bojar, Barry Haddow, Rico Sennrich, Frédéric Blain, Lucia Specia, Jan Niehues, Alex Waibel, Alexandre Allauzen, Lauriane Aufrant, Franck Burlot, Elena Knyazeva, Thomas Lavergne, François Yvon, Stella Frank, and Mārcis Pinnis. 2016. The QT21/HimL Combined Machine Translation System. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.

Holger Schwenk and Jean Senellart. 2009. Translation Model Adaptation for an Arabic/French News Translation System by Lightly-Supervised Training. In *Proc. of the MT Summit XII*, Ottawa, Canada, August.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 182–189, Waikiki, HI, USA, October.

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.

Francis M. Tyers and Murat Serdar Alperen. 2010. South-East European Times: A parallel corpus of Balkan languages. In *Proc. of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53, Malta, May.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2010. Jane: Open Source Hierarchical Translation, Extended with Reordering and Lexicon

Models. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 262–270, Uppsala, Sweden, July.

David Vilar, Daniel Stein, Matthias Huck, and Hermann Ney. 2012. Jane: an advanced freely available hierarchical machine translation toolkit. *Machine Translation*, 26(3):197–216, September.

Philip Williams, Rico Sennrich, Maria Nădejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's Statistical Machine Translation Systems for WMT16. In *Proc. of the ACL 2016 First Conf. on Machine Translation (WMT16)*, Berlin, Germany, August.