The Edinburgh Machine Translation Systems for IWSLT 2015

Matthias Huck, Alexandra Birch

School of Informatics
University of Edinburgh
Scotland, United Kingdom

mhuck@inf.ed.ac.uk a.birch@ed.ac.uk

Abstract

This paper describes the University of Edinburgh's machine translation (MT) systems for the IWSLT 2015 evaluation campaign. Our submissions are based on preliminary systems which are under development for the purpose of lecture translation in the TraMOOC project,[1] funded by the European Union.

We participated in the English→Chinese and the English→German translation tasks in the MT track, utilizing only data supplied by the organizers or listed as permissible. We built phrase-based translation systems for both tasks. For English→German, we furthermore made use of 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, 3] for the English→Chinese and English→German translation tasks, and additionally a string-to-tree syntax-based system [4, 5] for English→German. Our primary submission translations for English→Chinese are the output of a single phrase-based system, whereas our primary submission translations for English→German are the output of a system combination [6] of two phrase-based systems and one syntax-based system.

The setups for our phrase-based systems have evolved from the configurations of the engines we built for Edinburgh's participation in last year's IWSLT evaluation [7] and in this year's Workshop on Statistical Machine Translation (WMT) shared translation task [8].

Edinburgh's syntax-based systems have recently yielded state-of-the-art performance on English→German news translation tasks [9, 10] and have been applied in an IWSLT-style setting for the first time for our last year's contrastive submission [7]. This year, a syntax-based system became part of our primary submission by contributing input to a system combination.

For system combination, we employed the implementation that has been released as part of the Jane machine translation toolkit [11]. Multiple previous top-ranked submissions to open evaluation campaigns have relied on this system combination framework [12, 13, 14].

[1]http://tramooc.eu

2. System Overview

2.1. Training and Tuning

For both the phrase-based systems and the syntax-based system, we first preprocess the parallel training data and then create word alignments by aligning the data in both directions with MGIZA++ [15]. We use a sequence of IBM word alignment models [16] with five iterations of EM training [17] 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 [18, 3] to the two trained alignments. We extract bilingual phrases that are consistent with the symmetrized word alignment from the parallel training data. In the case of the syntax-based system, we also need syntactic parses of the target-language side of the parallel training data in order to extract synchronous context-free grammar rules.

We train n-gram language models (LMs) with modified Kneser-Ney smoothing [19, 20]. KenLM [21] is employed for LM training and scoring, and SRILM [22] for linear LM interpolation.

Our translation model incorporates a number of different features in a log-linear combination [23]. We tune the feature weights with batch k-best MIRA [24] to maximize BLEU [25] on a development set. We run MIRA for 25 iterations on 200-best lists (phrase-based) or 1000-best lists (syntax-based).

In our experiments (cf. Section 3) with the phrase-based system, we commence with a plain baseline which comprises a small amount of vital features only. We then incrementally extend the system with further features and more advanced techniques. Each setup is re-tuned individually to obtain optimal feature weights for the respective configuration.

2.2. Phrase-based System

The features of our plain phrase-based baseline are:

- Phrase translation log-probabilities in both target-to-source and source-to-target direction.
- Lexical translation log-probabilities in both target-to-source and source-to-target direction.
- Word penalty.
- Phrase penalty.
- A distance-based distortion cost.
- A 5-gram language model over words. Singleton n-grams of order three and higher are discarded.

We extract phrases up to a length of five. We prune the phrase table to a maximum of 100 best translation options per distinct source side and apply a minimum score threshold τ on the source-to-target phrase translation probability, with $\tau = 0.0001$ during tuning and $\tau = 0.00001$ during testing. We use cube pruning [26] in decoding. Pop limit and stack limit are set to 1000 for tuning and to 5000 for testing. A distortion limit of six is enforced during decoding, and we disallow reordering over punctuation. Furthermore, Minimum Bayes Risk decoding [27] is employed for testing.

Extensions we experimented with for either English→German or English→Chinese are:

LRM. A hierarchical lexicalized reordering model [28]. This model estimates the probabilities of orientation classes for each phrase from the training data. We use four orientation classes: *monotone*, *swap*, *left-discontinuous*, and *right-discontinuous*.

TM factors. Translation model (TM) factors beyond word surface forms [29, 30]. Factors can for instance be part-of-speech (POS) tag, morphological tag, or automatically learnt word classes, e.g. from `mkcls` [31]. Factors can be added on either source side or target side or both. We do not use a generation step but merely enrich the phrases with factored annotation. The annotation is obtained by tagging the training data prior to phrase extraction. Source-side factors such as POS or morphological tags can be helpful for disambiguating phrases: at decoding time, we annotate the input text in a preprocessing step, and the decoder only applies phrases with matching annotation. Target-side factors can be helpful for providing a longer context window via n-gram models of higher order over representations given by the factors (which we mention next in this list).

7-gram class-based LM. A 7-gram language model over `mkcls` word classes.

7-gram POS LM. A 7-gram language model over part-of-speech tags.

7-gram morph LM. A 7-gram language model over morphological tags.

Good-Turing smoothing. Good-Turing smoothing of phrase translation probabilities [32].

Count features. Seven binary features indicating absolute occurrence count classes of phrase pairs.

Sparse features. Sparse phrase length features, and sparse lexical features for the top 200 words.

Domain indicators. Binary features indicating the provenance of phrase pairs: if a phrase pair has been seen in a particular training corpus, a binary indicator associated with the respective training corpus fires on application of that phrase pair during decoding.

Phrase table fill-up. A foreground phrase table extracted from in-domain data is filled up with entries from a background phrase table extracted from all data [33, 34]. An entry from the background table is only added if the foreground table does not know the respective phrase identity. A binary feature distinguishes background phrases from foreground phrases. (The baseline uses a phrase table extracted from all data.)

5-gram OSM. A 5-gram operation sequence model [35].

5-gram OSM over word classes. A 5-gram operation sequence model over `mkcls` word classes.

5-gram OSMs over factors. Operation sequence models over various representations given by the factors.

In-domain OSMs. 5-gram operation sequence models over words and factors, trained on the in-domain portion of the parallel data only.

Unpruned LM. The baseline 5-gram language model over words is replaced by a version where singleton n-grams of order three and higher have not been discarded.

No singleton phrases. Phrase pairs with an occurrence count of one are removed from the phrase table.

Sparse LR. Sparse lexicalized reordering features [36] with weights learnt via RPROP with a maximum expected BLEU objective [37, 38]. The features are added on top of the standard hierarchical lexicalized reordering model. We apply features based on all words as well as word classes with 200 clusters on both source and target side. Active feature groups are *between*, *phrase*, and *stack*. We follow a similar training procedure as suggested by Wuebker et al. [38].[2] Maximum expected BLEU training with RPROP is conducted on the in-domain fraction of the training data. We train on 100-best lists. We set the regularization parameter to 10^{-5} and use the weights obtained after 50 iterations of RPROP. Rather than decoding the training data with leaving-one-out, we utilize a system with no singleton phrases. The learnt sparse lexicalized reordering features are condensed to a single feature per orientation, as suggested by Auli et al. [37]. A final MIRA run tunes weights for those condensed features along with the other features in the log-linear model of the translation system.

[2]Our tool for maximum expected BLEU training has been released as part of the Moses code base on GitHub.

2.3. Syntax-based System

The syntactic translation model for our string-to-tree system conforms to the GHKM syntax approach as proposed by **G**alley, **H**opkins, **K**night, and **M**arcu [4] with composed rules [39, 40]. Decoding is carried out with a procedure based on bottom-up chart parsing. The parsing algorithm is extended to handle translation candidates and to incorporate language model scores via cube pruning [26].

Standard features of Edinburgh's string-to-tree syntax-based systems are:

- Rule translation log-probabilities in both target-to-source and source-to-target direction, smoothed with Good-Turing discounting.
- Lexical translation log-probabilities in both target-to-source and source-to-target direction.
- Word penalty.
- Rule penalty.
- A rule rareness penalty.
- The monolingual PCFG probability of the tree fragment from which the rule was extracted.
- A 5-gram language model over words.

When extracting syntactic rules, we impose several restrictions for composed rules, in particular a maximum number of 100 tree nodes per rule, a maximum depth of seven, and a maximum size of seven. We discard rules with non-terminals on their right-hand side if they are singletons in the training data. Only the 200 best translation options per distinct rule source side with respect to the weighted rule-level model scores are loaded by the decoder. Search is carried out with a maximum chart span of 25, a rule limit of 500, a stack limit of 200, and a pop limit of 1000 for cube pruning [41]. During tuning, we constrain the translation options per rule source side to the top 20 candidates for faster optimization, and we set the cube pruning pop limit to 500. We configure Moses' `n-best-factor` parameter at a value of 100 to avoid short *n*-best lists.

For our IWSLT English→German syntax-based system, the target side of the parallel training data is parsed with BitPar [42]. We remove grammatical case and function information from the annotation obtained with BitPar and apply right binarization of the German parse trees prior to rule extraction [43, 44, 45].

The system is adapted to the TED domain by extracting two separate rule tables (from in-domain data and from out-of-domain parallel data) and merging them with a fill-up technique [33]. We also integrate a second 5-gram LM trained on the in-domain corpus into the log-linear combination. Additionally we add soft source syntactic constraints [46] and augment the system with non-syntactic phrases [47].

2.4. System Combination

The Jane machine translation toolkit implements a system combination approach via confusion network decoding [11]. The hypotheses from individual MT systems are aligned to each other with METEOR [48]. A confusion network is generated which represents all combined translations that can be produced from the set of individual hypotheses. The optimal combined hypothesis is chosen by finding the best path through the confusion network. The decision process is guided by a couple of simple features:

- Binary system voting features.
- A primary system indicator.
- Word penalty.
- A small 3-gram language model trained only on the set of individual hypotheses.
- A conventional 5-gram language model.

Feature weights are optimized with MERT [49].

We combine three individual systems with this method for our English→German primary submission.

3. Experiments

3.1. English→German MT

For the English→German MT task, we submitted outputs of two different phrase-based systems (*contrastive 1* and *contrastive 2*), a syntax-based system (*contrastive 3*), and a system combination (*primary*) of those three single systems. Table 1 shows their respective performance in terms of BLEU scores, along with the official scores [50] of the best last year's submission for comparison.

Our English→German systems are trained using monolingual and parallel data from the in-domain WIT[3] corpus [52], as well as Europarl [53], MultiUN [54], the parallel corpus from the Wikipedia [55] as provided for the evaluation campaign, the German Political Speeches corpus [56], and the permissible corpora from the WMT shared translation task [57]. For the systems with factors, annotation exploited in addition to word surface forms is: part-of-speech tags [58] on the English side; morphological tags [59] and part-of-speech tags [59] on the German side; and word classes from `mkcls` with 50 clusters on both sides.

5-gram LMs over words are estimated over a concatenation of all target-language training data, rather than linearly interpolating individual LMs over the different corpora. We found this to perform equally well or better on the given task. Class-based LMs, POS LMs, and morph LMs, on the other hand, are linear interpolations of individually trained LMs.[3] Feature weights for all single engines are tuned on a concatenation of `TED.dev2010`, `TED.tst2010`, `TEDX.dev2012`, and `TEDX.tst2013`.[4]

[3]Individual LMs over factors are trained with KenLM's `--discount_fallback --prune '0 0 1'` parameters.

[4]Note that `TEDX.tst2013` and `tst2013` (= `TED.tst2013`) are two different sets.

en→de	tst2011	tst2012	tst2013	tst2014	tst2015
phrase-based (*contrastive 1*)	28.3	24.7	26.3	23.3	25.4
phrase-based w/o singleton phrases + sparse LR (*contrastive 2*)	27.9	24.5	26.8	23.3	25.5
syntax-based (*contrastive 3*)	26.8	23.6	26.1	22.7	24.3
system combination (*primary*)	28.4	25.6	27.0	24.0	26.0
best IWSLT 2014 submission (*EU-BRIDGE* [14])	–	–	26.2	23.3	–

Table 1: Edinburgh submission system results for the English→German MT task (case-sensitive BLEU scores), and results of the best IWSLT 2014 submission as reported by Cettolo et al. [50]. The Edinburgh *primary* submission is a system combination of the three *contrastive* systems and was tuned on tst2012.

en→zh	tst2012	tst2013	tst2014	tst2015
phrase-based (*primary*)	21.3	22.9	19.6	25.4
best IWSLT 2014 submission (*USTC* [51])	–	22.5	21.6	–

Table 2: Edinburgh submission system results for the English→Chinese MT task (character-based BLEU scores), and results of the best IWSLT 2014 submission as reported by Cettolo et al. [50].

Phrase-based system. Table 3 presents the results achieved with the plain phrase-based baseline, and the gains when incrementally adding extensions as described in Section 2.2.[5] The *contrastive 1* submission system outperforms the plain baseline by up to +3.6 BLEU points (on `tst2011`). If we remove singleton phrases on top of that, we observe a small gain on `tst2013`, but performance degrades slightly on `tst2011` and `tst2012`. The sparse lexicalized reordering features trained via RPROP with a maximum expected BLEU objective (*contrastive 2*) do not further affect the results too much.[6] However, the *contrastive 2* submission system outperforms the plain baseline by +3.5 BLEU points on a different test set (on `tst2013`).

Syntax-based system. In the syntax-based system, we utilize neither the parallel corpus from the Wikipedia nor MultiUN or the German Political Speeches corpus for rule extraction.[7] We only use the target side of the Wikipedia corpus as LM training data. The development set is the same as for the phrase-based systems. Our IWSLT string-to-tree syntax-based system (*contrastive 3*) is outperformed by the phrase-based submission systems by a bit more than one BLEU point on this year's evaluation set (`tst2015`), cf. Table 1. The average BLEU delta on the other test sets is lower, though.

System combination. The parameters of the system combination (*primary*) are optimized on `tst2012`. The consensus translation produced by the system combination boosts the BLEU score by half a point over the best single system on this year's evaluation set (`tst2015`), cf. Table 1. Improvements on the other test sets vary between +0.1 and

en→de	tst2011	tst2012	tst2013
phrase-based baseline	24.7	22.0	23.3
+ LRM	25.5	22.0	24.1
+ TM factors	25.3	22.1	23.8
+ 7-gram class-based LM	25.9	22.5	24.2
+ 7-gram POS LM	26.1	22.8	24.6
+ 7-gram morph LM	26.5	22.9	24.9
+ Good-Turing smoothing	26.8	23.6	24.9
+ count features	26.8	23.4	24.9
+ sparse features	26.9	23.7	25.1
+ domain indicators	27.2	23.6	25.3
+ 5-gram OSM	27.6	24.1	26.1
+ 5-gram OSMs over factors	27.8	24.3	26.0
+ in-domain OSMs	28.0	24.3	26.3
+ unpruned LM (*contrastive 1*)	28.3	24.7	26.3
+ no singleton phrases	27.9	24.6	26.7
+ sparse LR (*contrastive 2*)	27.9	24.5	26.8

Table 3: Incremental improvements over a plain phrase-based baseline for English→German (case-sensitive BLEU scores).

en→zh	tst2012	tst2013
phrase-based baseline	19.2	21.0
+ LRM	19.8	21.7
+ Good-Turing smoothing	20.0	21.9
+ count features	20.1	21.9
+ 7-gram class-based LM (in-domain)	20.0	22.0
+ phrase table fill-up	21.0	22.3
+ 5-gram OSM	21.0	22.5
+ 5-gram OSM over word classes	20.9	22.5
+ in-domain OSMs (*primary*)	21.3	22.9

Table 4: Incremental improvements over a plain phrase-based baseline for English→Chinese (character-based BLEU scores).

[5] The order in which extensions are added is not motivated by any specific rationale other than our personal preference.

[6] We add the *sparse LR* to the system without singleton phrases. This avoids a mismatch with the system used in n-best generation for maximum expected BLEU training.

[7] Due to time constraints, these corpora have been omitted for the benefit of faster training.

+0.7 (disregarding tst2012, since it has been used to tune the system combination).

Our best single system yields translation quality on the level of the last year's best submission, which was a system combination [14]. Our primary submission is around 0.7 BLEU points better than last year's best submission.

3.2. English→Chinese MT

For the English→Chinese MT task, we submitted the output of a phrase-based single system (*primary*). Table 2 shows the performance in terms of BLEU scores, measured on character level with the aid of the Chinese character tokenization script provided by the organizers of the evaluation campaign. For comparison, we also include the official scores [50] of the best last year's submission.

Our English→Chinese systems are trained using monolingual and parallel data from the in-domain WIT^3 corpus [52], as well as MultiUN [54]. For the English-Chinese MultiUN parallel data, we resorted to the sentence-aligned version as distributed in OPUS [60]. We perform Chinese word segmentation with the Stanford Word Segmenter [61] as a preprocessing step on all target-side data. The character-based tokenization is conducted for evaluation purposes only, whereas our models operate on word-segmented data.

Table 4 presents the results achieved with the plain phrase-based baseline, and the gains when incrementally adding extensions as described in Section 2.2. The 5-gram LM over words is a linear interpolation of individual LMs, the 7-gram class-based LM is trained on in-domain data only. The only factors we use for English→Chinese are word classes from mkcls with 50 clusters. Feature weights are tuned on a concatenation of dev2010, tst2010, and tst2011. The submission system outperforms the plain baseline by up to +2.1 BLEU points (on tst2012).

The comparison with last year's best submission [51] is somewhat surprising: the BLEU score of our system is +0.4 points higher on tst2013, but we significantly lag behind on tst2014. We are currently unaware of the reason for this behavior.

4. Summary

We built high-quality machine translation systems for the IWSLT 2015 English→Chinese and English→German translation tasks in the MT track. By utilizing advanced features and techniques, we have been able to achieve improvements over plain phrase-based baselines of two BLEU points or more on both language pairs. All methods we employed are implemented in publicly available software such as the Moses and the Jane statistical machine translation toolkits.

5. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 644333 (*TraMOOC*).

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. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Prague, Czech Republic, June 2007, pp. 177–180.

[2] R. Zens, F. J. Och, and H. Ney, "Phrase-Based Statistical Machine Translation," in *German Conf. on Artificial Intelligence*, Aachen, Germany, Sept. 2002, pp. 18–32.

[3] P. Koehn, F. J. Och, and D. Marcu, "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)*, Edmonton, Canada, May/June 2003, pp. 127–133.

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

[5] P. Williams and P. Koehn, "GHKM Rule Extraction and Scope-3 Parsing in Moses," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Montréal, Canada, June 2012, pp. 388–394.

[6] E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System Combination for Machine Translation of Spoken and Written Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1222–1237, Sept. 2008.

[7] A. Birch, M. Huck, N. Durrani, N. Bogoychev, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, USA, Dec. 2014, pp. 49–56.

[8] B. Haddow, M. Huck, A. Birch, N. Bogoychev, and P. Koehn, "The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal, Sept. 2015, pp. 126–133.

[9] 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.

[10] P. Williams, R. Sennrich, M. Nadejde, M. Huck, and P. Koehn, "Edinburgh's Syntax-Based Systems at WMT 2015," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal, Sept. 2015, pp. 199–209.

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

[12] M. Freitag, S. Peitz, J. Wuebker, H. Ney, N. Durrani, M. Huck, P. Koehn, T.-L. Ha, J. Niehues, M. Mediani, T. Herrmann, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, Dec. 2013, pp. 128–135.

[13] M. Freitag, S. Peitz, J. Wuebker, H. Ney, M. Huck, R. Sennrich, N. Durrani, M. Nadejde, P. Williams, P. Koehn, T. Herrmann, E. Cho, and A. Waibel, "EU-BRIDGE MT: Combined Machine Translation," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June 2014, pp. 105–113.

[14] 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)*, Lake Tahoe, CA, USA, Dec. 2014, pp. 57–64.

[15] 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.

[16] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, June 1993.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, 1977.

[18] 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.

[19] R. Kneser and H. Ney, "Improved Backing-Off for M-gram Language Modeling," in *Proceedings of the Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Detroit, MI, USA, May 1995, pp. 181–184.

[20] 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.

[21] K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Edinburgh, Scotland, UK, July 2011, pp. 187–197.

[22] A. Stolcke, "SRILM – an Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, vol. 3, Denver, CO, USA, Sept. 2002.

[23] 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.

[24] C. Cherry and G. Foster, "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)*, Montréal, Canada, June 2012, pp. 427–436.

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

[26] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, June 2007.

[27] S. Kumar and W. Byrne, "Minimum Bayes-Risk Decoding 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)*, Boston, MA, USA, May 2004, pp. 169–176.

[28] 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.

[29] P. Koehn and H. Hoang, "Factored Translation Models," in *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and*

Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, June 2007, pp. 868–876.

[30] 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.

[31] F. J. Och, "An Efficient Method for Determining Bilingual Word Classes," in *Proc. of the Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, Bergen, Norway, June 1999, pp. 71–76.

[32] G. Foster, R. Kuhn, and H. Johnson, "Phrasetable Smoothing for Statistical Machine Translation," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Sydney, Australia, July 2006, pp. 53–61.

[33] 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.

[34] J. Niehues and A. Waibel, "Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT," in *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, Oct./Nov. 2012.

[35] N. Durrani, A. Fraser, and H. Schmid, "Model With Minimal Translation Units, But Decode With Phrases," in *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA, USA, June 2013, pp. 1–11.

[36] C. Cherry, "Improved Reordering for Phrase-Based Translation using Sparse Features," in *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA, USA, June 2013, pp. 22–31.

[37] M. Auli, M. Galley, and J. Gao, "Large-scale Expected BLEU Training of Phrase-based Reordering Models," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1250–1260.

[38] J. Wuebker, S. Muehr, P. Lehnen, S. Peitz, and H. Ney, "A Comparison of Update Strategies for Large-Scale Maximum Expected BLEU Training," in *Proc. of the Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Denver, CO, USA, May 2015, pp. 1516–1526.

[39] M. Galley, J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer, "Scalable Inference and Training of Context-Rich Syntactic Translation Models," in *Proc. of the 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics*, Sydney, Australia, July 2006, pp. 961–968.

[40] S. DeNeefe, K. Knight, W. Wang, and D. Marcu, "What Can Syntax-Based MT Learn from Phrase-Based MT?" in *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, June 2007, pp. 755–763.

[41] M. Huck, D. Vilar, M. Freitag, and H. Ney, "A Performance Study of Cube Pruning for Large-Scale Hierarchical Machine Translation," in *Proceedings of the NAACL 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Atlanta, GA, USA, June 2013, pp. 29–38.

[42] 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.

[43] W. Wang, K. Knight, and D. Marcu, "Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy," in *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, June 2007, pp. 746–754.

[44] W. Wang, J. May, K. Knight, and D. Marcu, "Restructuring, Re-labeling, and Re-aligning for Syntax-based Machine Translation," *Computational Linguistics*, vol. 36, no. 2, pp. 247–277, June 2010.

[45] 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.

[46] M. Huck, H. Hoang, and P. Koehn, "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.

[47] ——, "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.

[48] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of

Machine Translation Systems," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Edinburgh, Scotland, UK, July 2011, pp. 85–91.

[49] F. J. Och, "Minimum Error Rate Training for Statistical Machine Translation," in *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[50] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, USA, Dec. 2014, pp. 2–17.

[51] S. Wang, Y. Wang, J. Li, Y. Cui, and L. Dai, "The USTC Machine Translation System for IWSLT 2014," in *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, USA, Dec. 2014, pp. 134–138.

[52] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

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

[54] A. Eisele and Y. Chen, "MultiUN: A Multilingual Corpus from United Nation Documents," in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, May 2010, pp. 2868–2872.

[55] K. Wołk and K. Marasek, "Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs," *Procedia Technology*, vol. 18, pp. 126–132, 2014, International workshop on Innovations in Information and Communication Science and Technology (IICST), 3-5 September 2014, Warsaw, Poland.

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

[57] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi, "Findings of the 2015 Workshop on Statistical Machine Translation," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal, September 2015, pp. 1–46.

[58] A. Ratnaparkhi, "A Maximum Entropy Part-Of-Speech Tagger," in *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Philadelphia, PA, USA, May 1996.

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

[60] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proc. of the Int. Conf. on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, May 2012, pp. 2214–2218.

[61] P.-C. Chang, M. Galley, and C. Manning, "Optimizing Chinese Word Segmentation for Machine Translation Performance," in *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Columbus, OH, USA, June 2008, pp. 224–232.