Neural Machine Translation by Minimising the Bayes-risk with Respect to Syntactic Translation Lattices

Felix Stahlberg[1], **Adrià de Gispert**[1,2], **Eva Hasler**[1,2] **and Bill Byrne**[1,2]

[1]Department of Engineering, University of Cambridge, UK
[2]SDL Research, Cambridge, UK
{fs439,ad465,ech57,wjb31}@cam.ac.uk
{agispert,ehasler,bbyrne}@sdl.com

Abstract

We present a novel scheme to combine neural machine translation (NMT) with traditional statistical machine translation (SMT). Our approach borrows ideas from linearised lattice minimum Bayes-risk decoding for SMT. The NMT score is combined with the Bayes-risk of the translation according the SMT lattice. This makes our approach much more flexible than n-best list or lattice rescoring as the neural decoder is not restricted to the SMT search space. We show an efficient and simple way to integrate risk estimation into the NMT decoder which is suitable for word-level as well as subword-unit-level NMT. We test our method on English-German and Japanese-English and report significant gains over lattice rescoring on several data sets for both single and ensembled NMT. The MBR decoder produces entirely new hypotheses far beyond simply rescoring the SMT search space or fixing UNKs in the NMT output.

1 Introduction

Lattice minimum Bayes-risk (LMBR) decoding has been applied successfully to translation lattices in traditional SMT to improve translation performance of a single system (Kumar and Byrne, 2004; Tromble et al., 2008; Blackwood et al., 2010). However, minimum Bayes-risk (MBR) decoding is also a very powerful framework for combining diverse systems (Sim et al., 2007; de Gispert et al., 2009). Therefore, we study combining traditional SMT and NMT in a hybrid decoding scheme based on MBR. We argue that MBR-based methods in their present form are not well-suited for NMT because of the following reasons:

- Previous approaches work well with rich lattices and diverse hypotheses. However, NMT decoding usually relies on beam search with a limited beam and thus produces very narrow lattices (Li and Jurafsky, 2016; Vijayakumar et al., 2016).

- NMT decoding is computationally expensive. Therefore, it is difficult to collect the statistics needed for risk calculation for NMT.

- The Bayes-risk in SMT is usually defined for complete translations. Therefore, the risk computation needs to be restructured in order to integrate it in an NMT decoder which builds up hypotheses from left to right.

To address these challenges, we use a special loss function which is computationally tractable as it avoids using NMT scores for risk calculation. We show how to reformulate the original LMBR decision rule for using it in a word-based NMT decoder which is not restricted to an n-best list or a lattice. Our hybrid system outperforms lattice rescoring on multiple data sets for English-German and Japanese-English. We report similar gains from applying our method to subword-unit-based NMT rather than word-based NMT.

2 Combining NMT and SMT by Minimising the Lattice Bayes-risk

We propose to collect statistics for MBR from a potentially large translation lattice generated with SMT, and use the n-gram posteriors as additional score in NMT decoding. The LMBR decision rule used by Tromble et al. (2008) has the form

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 362–368,
Valencia, Spain, April 3-7, 2017.

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_h} \Big(\underbrace{\Theta_0 |\mathbf{y}| + \sum_{\mathbf{u} \in \mathcal{N}} \Theta_{|\mathbf{u}|} \#_{\mathbf{u}}(\mathbf{y}) P(\mathbf{u}|\mathcal{Y}_e)}_{:= E(\mathbf{y})} \Big) \quad (1)$$

where $\mathcal{Y}_h$ is the *hypothesis space* of possible translations, $\mathcal{Y}_e$ is the *evidence space* for computing the Bayes-risk, and $\mathcal{N}$ is the set of all n-grams in $\mathcal{Y}_e$ (typically, $n = 1 \dots 4$). In this work, our evidence space $\mathcal{Y}_e$ is a translation lattice generated with SMT. The function $\#_{\mathbf{u}}(\mathbf{y})$ counts how often n-gram $\mathbf{u}$ occurs in translation $\mathbf{y}$. $P(\mathbf{u}|\mathcal{Y}_e)$ denotes the path posterior probability of $\mathbf{u}$ in $\mathcal{Y}_e$. Our aim is to integrate these n-gram posteriors into the NMT decoder since they correlate well with the presence of n-grams in reference translations (de Gispert et al., 2013). We call the quantity to be maximised the *evidence* $E(\mathbf{y})$ which corresponds to the (negative) Bayes-risk which is normally minimised in MBR decoding. We emphasize that this risk can be computed for any translation hypothesis and not only those produced by the SMT system.

NMT assigns a probability to a translation $\mathbf{y} = y_1^T$ of source sentence $\mathbf{x}$ via a left-to-right factorisation based on the chain rule:

$$P_{NMT}(y_1^T|\mathbf{x}) = \prod_{t=1}^{T} \underbrace{P_{NMT}(y_t|y_1^{t-1}, \mathbf{x})}_{= g(y_{t-1}, s_t, c_t)} \quad (2)$$

where $g(\cdot)$ is a neural network using the hidden state of the decoder network s_t and the context vector c_t which encodes relevant parts of the source sentence (Bahdanau et al., 2015).[1] $P_{NMT}(\cdot)$ can also represent an ensemble of NMT systems in which case the scores of the individual systems are multiplied together to form a single distribution. Applying the LMBR decision rule in Eq. 1 directly to NMT would involve computing $P_{NMT}(\mathbf{y}|\mathbf{x})$ for all translations in the evidence space. In case of LMBR this is equivalent to rescoring the entire translation lattice exhaustively with NMT. However, this is not feasible even for small lattices because the evaluation of $g(\cdot)$ is computationally very expensive. Therefore, we propose to calculate the Bayes-risk over SMT translation lattices using only pure SMT scores, and bias the NMT decoder towards low-risk hypotheses. Our final combined decision rule is

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \Big(E(\mathbf{y}) + \lambda \log P_{NMT}(\mathbf{y}|\mathbf{x}) \Big). \quad (3)$$

If $\mathbf{y}$ contains a word not in the NMT vocabulary, the NMT model provides a score and updates its decoder state as for an unknown word. We note that $E(\mathbf{y})$ can be computed even if $\mathbf{y}$ is not in the SMT lattice. Therefore, Eq. 3 can be used to generate translations outside the SMT search space. We further note that Eq. 3 can be derived as an instance of LMBR under a modified loss function.

3 Left-to-right Decoding

Beam search is often used for NMT because the factorisation in Eq. 2 allows to build up hypotheses from left to right. In contrast, our definition of the *evidence* in Eq. 1 contains a sum over the (unordered) set of all n-grams. However, we can rewrite our objective function in Eq. 3 in a way which makes it easy to use with beam search.

$$\begin{aligned} & E(\mathbf{y}) + \lambda \log P_{NMT}(\mathbf{y}|\mathbf{x}) \\ =& \Theta_0 |\mathbf{y}| + \sum_{\mathbf{u} \in \mathcal{N}} \Theta_{|\mathbf{u}|} \#_{\mathbf{u}}(\mathbf{y}) P(\mathbf{u}|\mathcal{Y}_e) \\ & + \lambda \sum_{t=1}^{T} \log P_{NMT}(y_t|y_1^{t-1}, \mathbf{x}) \\ =& \sum_{t=1}^{T} \Big(\Theta_0 + \sum_{n=1}^{4} \Theta_n P(y_{t-n}^t|\mathcal{Y}_e) \\ & + \lambda \log P_{NMT}(y_t|y_1^{t-1}, \mathbf{x}) \Big) \end{aligned} \quad (4)$$

for n-grams up to order 4. This form lends itself naturally to beam search: at each time step, we add to the previous partial hypothesis score both the log-likelihood of the last token according the NMT model, and the partial MBR gains from the current n-gram history. Note that this is similar to applying (the exponentiated scores of) an interpolated language model based on n-gram posteriors extracted from the SMT lattice. In the remainder of this paper, we will refer to decoding according Eq. 4 as *MBR-based* NMT.

4 Efficient n-gram Posterior Calculation

The risk computation in our approach is based on posterior probabilities $P(\mathbf{u}|\mathcal{Y}_e)$ for n-grams $\mathbf{u}$

[1] We refer to Bahdanau et al. (2015) for a full discussion of attention-based NMT.

Setup		news-test2014	news-test2015	news-test2016
SMT baseline (de Gispert et al., 2010, HiFST)		18.9	21.2	26.0
Single NMT (word)	Pure NMT	17.7	19.6	23.1
	100-best rescoring	20.6	22.5	27.5
	Lattice rescoring	21.6	23.8	**29.6**
	This work	**22.0**	**24.6**	29.5
5-Ensemble NMT (word)	Pure NMT	19.4	21.8	25.4
	100-best rescoring	21.0	23.3	28.6
	Lattice rescoring	22.1	24.2	30.2
	This work	**22.8**	**25.4**	**30.8**
Single NMT (BPE)	Pure NMT	19.6	21.9	24.6
	Lattice rescoring	21.5	24.0	**29.6**
	This work	**21.7**	**24.1**	28.6
3-Ensemble NMT (BPE)	Pure NMT	21.0	23.4	27.0
	Lattice rescoring	21.7	24.2	**30.0**
	This work	**22.3**	**24.9**	29.2

Table 1: English-German lower-cased BLEU scores calculated with `mteval-v13a.pl`.[2]

which we extract from the SMT translation lattice $\mathcal{Y}_e$. $P(\mathbf{u}|\mathcal{Y}_e)$ is defined as the sum of the path probabilities $P_{SMT}(\cdot)$ of paths in $\mathcal{Y}_e$ containing $\mathbf{u}$ (Blackwood et al., 2010, Eq. 2):

$$P(\mathbf{u}|\mathcal{Y}_e) = \sum_{\mathbf{y} \in \{\mathbf{y} \in \mathcal{Y}_e : \#_{\mathbf{u}}(\mathbf{y}) > 0\}} P_{SMT}(\mathbf{y}|\mathbf{x}). \quad (5)$$

We use the framework of Blackwood et al. (2010) based on n-gram mapping and path counting transducers to efficiently pre-compute all non-zero values of $P(\mathbf{u}|\mathcal{Y}_e)$. Complete enumeration of all n-grams in a lattice is usually feasible even for very large lattices (Blackwood et al., 2010). Additionally, for all these n-grams $\mathbf{u}$, we smooth $P(\mathbf{u}|\mathcal{Y}_e)$ by mixing it with the uniform distribution to flatten the distribution and increase the offset to n-grams which are not in the lattice.

5 Subword-unit-based NMT

Character-based or subword-unit-based NMT (Chitnis and DeNero, 2015; Sennrich et al., 2016; Chung et al., 2016; Luong and Manning, 2016; Costa-Jussà and Fonollosa, 2016; Ling et al., 2015; Wu et al., 2016) does not use isolated words as modelling units but applies a finer grained tokenization scheme. One of the main motivation for these approaches is to overcome the limited vocabulary in word-based NMT. We consider our hybrid system as an alternative way to fix NMT OOVs. However, our method can also be used with subword-unit-based NMT. In this work, we use byte pair encodings (Sennrich et al., 2016, BPE) to test combining word-based SMT with subword-unit-based NMT via both lattice rescoring and MBR. First, we construct a finite state transducer (FST) which maps word sequences to BPE sequences. Then, we convert the word-based SMT lattices to BPE-based lattices by composing them with the mapping transducer and projecting the output tape using standard OpenFST operations (Allauzen et al., 2007). The converted lattices are used for extracting n-gram posteriors as described in the previous sections. Note that even though the n-grams are on the BPE level, their posteriors are computed from word-level SMT translation scores.

6 Experimental Setup

We test our approach on English-German (En-De) and Japanese-English (Ja-En). For En-De, we use the WMT *news-test2014* (the filtered version) as a development set, and keep *news-test2015* and *news-test2016* as test sets. For Ja-En, we use the ASPEC corpus (Nakazawa et al., 2016) to be strictly comparable to the evaluation done in the Workshop of Asian Translation (WAT).

The NMT systems are as described by Stahlberg et al. (2016b) using the Blocks and Theano frameworks (van Merriënboer et al., 2015; Bastien et al., 2012) with hyper-parameters as in (Bahdanau et al., 2015) and a vocabulary size of 30k for Ja-En and 50k for En-De. We use the coverage penalty proposed by Wu et al. (2016) to improve the length and coverage of translations. Our final ensembles combine five (En-De) to six (Ja-En) independently trained NMT systems.

Our En-De SMT baseline is a hierarchical system based on the HiFST package[3] which produces rich output lattices. The system uses rules ex-

[2]Comparable to `http://matrix.statmt.org/`

[3]`http://ucam-smt.github.io/`

Setup		dev	test
SMT baseline (Neubig, 2013, Travatar)		19.5	22.2
Single NMT (word)	Pure NMT	20.3	22.5
	10k-best rescoring	22.2	24.5
	This work	**22.4**	**25.2**
6-Ensemble NMT (word)	Pure NMT	22.6	25.0
	10k-best rescoring	22.4	25.4
	This work	**23.9**	**26.5**
Single NMT (BPE)	Pure NMT	20.8	23.5
	10k-best rescoring	21.9	24.6
	This work	**23.0**	**25.4**
3-Ensemble NMT (BPE)	Pure NMT	23.3	25.9
	10k-best rescoring	22.6	25.1
	This work	**24.1**	**26.7**

Table 2: Japanese-English cased BLEU scores calculated with Moses' `multi-bleu.pl`.[5]

tracted as described by de Gispert et al. (2010) and a 5-gram language model (Heafield et al., 2013).

In Ja-En we use Travatar (Neubig, 2013), an open-source tree-to-string system. We provide the system with Japanese trees obtained using the Ckylark parser (Oda et al., 2015) and train it on high-quality alignments as recommended by Neubig and Due (2014). This system, which reproduces the results of the best submission in WAT 2014 (Neubig, 2014), is used to create a 10k-best list of hypotheses, which we convert into determinised and minimised FSAs for our work. Our Ja-En NMT models are trained on the same 500k training samples as the Travatar baseline.

The parameter λ is tuned by optimising the BLEU score on the validation set, and we set $\Theta_i = 1$ $(i = 0, \ldots, 4)$. Using the BOBYQA algorithm (Powell, 2009) or lattice MERT (Macherey et al., 2008) to optimise the Θ-parameters independently did not yield improvements. The beam search implementation of the SGNMT decoder[4] (Stahlberg et al., 2016b) is used in all our experiments. We set the beam size to 20 for En-De and 12 for Ja-En.

7 Results

Our results are summarised in Tab. 1 and 2.[6] Our approach outperforms both single NMT and SMT baselines by up to 3.4 BLEU for En-De and 2.8 BLEU for Ja-En. Ensembling yields further gains across all test sets both for the NMT baselines and our MBR-based hybrid systems. We see substantial gains from our MBR-based method over lattice rescoring for both single and ensembled NMT on all test sets and language pairs except En-De *news-test2016*. On Ja-En, we report 26.7 BLEU[5], second to only one system (as of February 2017) that uses a number of techniques such as minimum risk training and a much larger vocabulary size which could also be used in our framework.

Our word-level NMT baselines suffer from their limited vocabulary since we do not apply post-processing techniques like UNK-replace (Luong et al., 2015). Therefore, NMT with subword units (BPE) consistently outperforms them by a large margin. Lattice rescoring and MBR yield large gains for both BPE-based and word-based NMT. However, the performance difference between BPE- and word-level NMT diminishes with lattice rescoring and MBR decoding: rescoring with NMT often performs on the same level for both words and subword units, and MBR-based NMT is often even better with a word-level NMT baseline. This indicates that subword units are often not necessary when the hybrid system has access to a large word-level vocabulary like the SMT vocabulary.

Note that the BPE lattice rescoring system is constrained to produce words in the output vocabulary of the syntactic SMT system and is prevented from inventing new target language words out of combinations of subword units. MBR imposes a soft version of such a constraint by biasing the BPE-based system towards words in the SMT search space.

The hypotheses produced by our MBR-based method often differ from the translations in the baseline systems. For example, 77.8% of the translations from our best MBR-based system on Ja-En cannot be found in the SMT 10k-best list,

[4] `http://ucam-smt.github.io/sgnmt/html/`

[5] Comparable to `http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/list.php?t=2`

[6] Instructions for reproducing our key results will be available upon publication at `http://ucam-smt.github.io/sgnmt/html/tutorial.html`

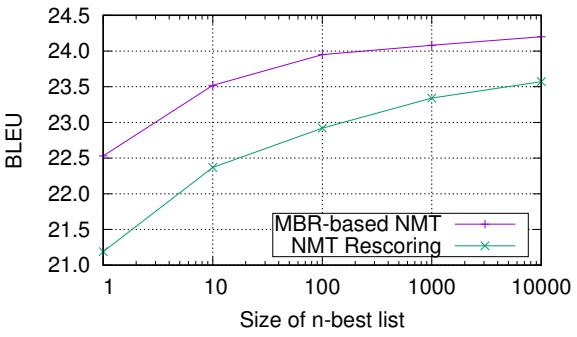

Figure 1: Performance over n-best list size on English-German *news-test2015*.

and 78.0% do not match the translation from the pure NMT 6-ensemble.[7] This suggests that our MBR decoder is able to produce entirely new hypotheses, and that our method has a profound effect on the translations which goes beyond rescoring the SMT search space or fixing UNKs in the NMT output.

Tab. 1 also shows that rescoring is sensitive to the size of the n-best list or lattice: rescoring the entire lattice instead of a 100-best list often yields a gain of 1 full BLEU point. In order to test our MBR-based method on small lattices, we compiled n-best lists of varying sizes to lattices and extracted n-gram posteriors from the reduced lattices. Fig. 1 shows that the n-best list size has an impact on both methods. Rescoring a 10-best list already yields a large improvement of 1.2 BLEU. However, the hypotheses are still close to the SMT baseline. The MBR-based approach can make better use of small n-best lists as it does not suffer this restriction. MBR-based combination on a 10-best list performs on about the same level as rescoring a 10,000-best list which demonstrates a practical advantage of MBR over rescoring.

8 Related Work

Combining the advantages of NMT and traditional SMT has received some attention in current research. A recent line of research attempts to integrate SMT-style translation tables into the NMT system (Zhang and Zong, 2016; Arthur et al., 2016; He et al., 2016). Wang et al. (2016) interpolated NMT posteriors with word recommendations from SMT and jointly trained NMT together with a gating function which assigns the weight between SMT and NMT scores dynamically. Neubig et al. (2015) rescored n-best lists from a syntax-based SMT system with NMT. Stahlberg et al. (2016b) restricted the NMT search space to a Hiero lattice and reported improvements over n-best list rescoring. Stahlberg et al. (2016a) combined Hiero and NMT via a loose coupling scheme based on composition of finite state transducers and translation lattices which takes the edit distance between translations into account. Our approach is similar to the latter one since it allows to divert from SMT and generate translations without derivations in the SMT system. This ability is crucial for NMT ensembles because SMT lattices are often too narrow for the NMT decoder (Stahlberg et al., 2016a). However, the method proposed by Stahlberg et al. (2016a) insists on a monotone alignment between SMT and NMT translations to calculate the edit distance. This can be computationally expensive and not appropriate for MT where word reorderings are common. The MBR decoding described here does not have this shortcoming.

9 Conclusion

This paper discussed a novel method for blending NMT with traditional SMT by biasing NMT scores towards translations with low Bayes-risk with respect to the SMT lattice. We reported significant improvements of the new method over lattice rescoring on Japanese-English and English-German and showed that it can make good use even of very small lattices and n-best lists.

In this work, we calculated the Bayes-risk over non-neural SMT lattices. In the future, we are planning to introduce neural models to the risk estimation while keeping the computational complexity under control, e.g. by using neural n-gram language models (Bengio et al., 2003; Vaswani et al., 2013) or approximations of NMT scores (Lecorvé and Motlicek, 2012; Liu et al., 2016) for n-gram posterior calculation.

Acknowledgments

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC grant EP/L027623/1).

We thank Graham Neubig for providing pretrained parsing and alignment models, as well as scripts, to allow perfect reproduction of the NAIST WAT 2014 submission.

[7] Up to NMT OOVs.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata*, pages 11–23, Berlin, Heidelberg. Springer.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *EMNLP*, pages 1557–1567, Austin, Texas, USA.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*, Toulon, France.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. In *NIPS*, South Lake Tahoe, Nevada, USA.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3:1137–1155.

Graeme Blackwood, Adrià de Gispert, and William Byrne. 2010. Efficient path counting transducers for minimum Bayes-risk decoding of statistical machine translation lattices. In *ACL*, pages 27–32, Uppsala, Sweden.

Rohan Chitnis and John DeNero. 2015. Variable-length word encodings for neural translation models. In *EMNLP*, pages 2088–2093, Lisbon, Portugal.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *ACL*, pages 1693–1703, Berlin, Germany.

Marta R. Costa-Jussà and José AR. Fonollosa. 2016. Character-based neural machine translation. In *ACL*, pages 357–361, Berlin, Germany.

Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *HLT-NAACL*, pages 73–76, Boulder, Colorado, USA.

Adrià de Gispert, Gonzalo Iglesias, Graeme Blackwood, Eduardo R Banga, and William Byrne. 2010. Hierarchical phrase-based translation with weighted finite-state transducers and shallow-n grammars. *Computational Linguistics*, 36(3):505–533.

Adrià de Gispert, Graeme Blackwood, Gonzalo Iglesias, and William Byrne. 2013. N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27(2):85–114.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with SMT features. In *AAAI*, pages 151–157, Phoenix, Arizona.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *ACL*, pages 690–696, Sofia, Bulgaria.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, pages 169–176, Boston, MA, USA.

Gwénolé Lecorvé and Petr Motlicek. 2012. Conversion of recurrent neural network language models to weighted finite state transducers for automatic speech recognition. Technical report, Idiap.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Xunying Liu, Xie Chen, Yongqiang Wang, Mark JF. Gales, and Philip C. Woodland. 2016. Two efficient lattice rescoring methods using recurrent neural network language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1438–1449.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *ACL*, pages 1054–1063, Berlin, Germany.

Minh-Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *ACL*, pages 11–19, Beijing, China.

Wolfgang Macherey, Franz Josef Och, Ignacio Thayer, and Jakob Uszkoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In *EMNLP*, pages 725–734, Honolulu, HI, USA.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. In *LREC*, pages 2204–2208, Portoroz, Slovenia.

Graham Neubig and Kevin Duh. 2014. On the elements of an accurate tree-to-string machine translation system. In *ACL*, pages 143–149, Baltimore, USA.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: NAIST at WAT2015. In *WAT*, Kyoto, Japan.

Graham Neubig. 2013. Travatar: A forest-to-string machine translation engine based on tree transducers. In *ACL*, pages 91–96, Sofia, Bulgaria.

Graham Neubig. 2014. Forest-to-string SMT for Asian language translation: NAIST at WAT 2014. In *WAT*, Kyoto, Japan.

Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. Ckylark: A more robust PCFG-LA parser. In *NAACL*, pages 41–45, Denver, Colorado, USA.

Michael JD. Powell. 2009. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*, pages 1715–1725, Berlin, Germany.

Khe Chai Sim, William J. Byrne, Mark JF. Gales, Hichem Sahbi, and Philip C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *ICASSP*, pages IV–105–IV–108, Honolulu, HI, USA. IEEE.

Felix Stahlberg, Eva Hasler, and Bill Byrne. 2016a. The edit distance transducer in action: The University of Cambridge English-German system at WMT16. In *WMT*, pages 377–384, Berlin, Germany.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016b. Syntactically guided neural machine translation. In *ACL*, pages 299–305, Berlin, Germany.

Roy W. Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum Bayes-risk decoding for statistical machine translation. In *EMNLP*, pages 620–629, Honolulu, HI, USA.

Bart van Merriënboer, Dzmitry Bahdanau, Vincent Dumoulin, Dmitriy Serdyuk, David Warde-Farley, Jan Chorowski, and Yoshua Bengio. 2015. Blocks and fuel: Frameworks for deep learning. *CoRR*.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *EMNLP*, pages 1387–1392, Seattle, USA.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Xing Wang, Zhengdong Lu, Zhaopeng Tu, Hang Li, Deyi Xiong, and Min Zhang. 2016. Neural machine translation advised by statistical machine translation. *CoRR*, abs/1610.05150.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jiajun Zhang and Chengqing Zong. 2016. Bridging neural machine translation and bilingual dictionaries. *arXiv preprint arXiv:1610.07272*.