SuperNMT: Neural Machine Translation with Semantic Supersenses and Syntactic Supertags

Eva Vanmassenhove
Adapt Centre
Dublin City University
Ireland
eva.vanmassenhove@adaptcentre.ie

Andy Way
Adapt Centre
Dublin City University
Ireland
andy.way@adaptcentre.ie

Abstract

In this paper we incorporate semantic supersensetags and syntactic supertag features into EN–FR and EN–DE factored NMT systems. In experiments on various test sets, we observe that such features (and particularly when combined) help the NMT model training to converge faster and improve the model quality according to the BLEU scores.

1 Introduction

Neural Machine Translation (NMT) models have recently become the state-of-the art in the field of Machine Translation (Bahdanau et al., 2014; Cho et al., 2014; Kalchbrenner et al., 2014; Sutskever et al., 2014). Compared to Statistical Machine Translation (SMT), the previous state-of-the-art, NMT performs particularly well when it comes to word-reorderings and translations involving morphologically rich languages (Bentivogli et al., 2016). Although NMT seems to partially 'learn' or generalize some patterns related to syntax from the raw, sentence-aligned parallel data, more complex phenomena (e.g. prepositional-phrase attachment) remain problematic (Bentivogli et al., 2016). More recent work showed that explicitly (Sennrich and Haddow, 2016; Nadejde et al., 2017; Bastings et al., 2017; Aharoni and Goldberg, 2017) or implicitly (Eriguchi et al., 2017) modeling extra syntactic information into an NMT system on the source (and/or target) side could lead to improvements in translation quality.

When integrating linguistic information into an MT system, following the central role assigned to syntax by many linguists, the focus has been mainly on the integration of syntactic features. Although there has been some work on semantic features for SMT (Banchs and Costa-Jussà, 2011), so far, no work has been done on enriching NMT systems with more general semantic features at the word-level. This might be explained by the fact that NMT models already have means of learning semantic similarities through word-embeddings, where words are represented in a common vector space (Mikolov et al., 2013). However, making some level of semantics more explicitly available at the word level can provide the translation system with a higher level of abstraction beneficial to learn more complex constructions. Furthermore, a combination of both syntactic and semantic features would provide the NMT system with a way of learning semantico-syntactic patterns.

To apply semantic abstractions at the word-level that enable a characterisation beyond that what can be superficially derived, coarse-grained semantic classes can be used. Inspired by Named Entity Recognition which provides such abstractions for a limited set of words, supersense-tagging uses an inventory of more general semantic classes for domain-independent settings (Schneider and Smith, 2015). We investigate the effect of integrating supersense features (26 for nouns, 15 for verbs) into an NMT system. To obtain these features, we used the *AMALGrAM 2.0* tool (Schneider et al., 2014; Schneider and Smith, 2015) which analyses the input sentence for Multi-Word Expressions as well as noun and verb supersenses. The features are integrated using the framework of Sennrich et al. (2016), replicating the tags for every subword unit obtained by byte-pair encoding (BPE). We further experiment with a combination of semantic supersenses and syntactic supertag features (CCG syntactic categories (Steedman, 2000) using EasySRL (Lewis et al., 2015)) and less complex features such as POS-tags, assuming that supersense-tags have the potential to be useful especially in combination with syntactic information.

Proceedings of ACL 2018, Student Research Workshop, pages 67–73
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

The remainder of this paper is structured as follows: First, in Section 2, the related work is discussed. Next, Section 3 presents the semantic and syntactic features used. The experimental set-up is described in Section 4 followed by the results in Section 5. Finally, We conclude and present some of the ideas for future work in Section 6.

2 Related Work

In SMT, various linguistic features such as stems (Toutanova et al., 2008) lemmas (Mareček et al., 2011; Fraser et al., 2012), POS-tags (Avramidis and Koehn, 2008), dependency labels (Avramidis and Koehn, 2008) and supertags (Hassan et al., 2007; Haque et al., 2009) are integrated using pre- or post-processing techniques often involving factored phrase-based models (Koehn and Hoang, 2007). Compared to factored NMT models, factored SMT models have some disadvantages: (a) adding factors increases the sparsity of the models, (b) the *n*-grams limit the size of context that is taken into account, and (c) features are assumed to be independent of each other. However, adding syntactic features to SMT systems led to improvements with respect to word order and morphological agreement (Williams and Koehn, 2012; Sennrich, 2015).

One of the main strengths of NMT is its strong ability to generalize. The integration of linguistic features can be handled in a flexible way without creating sparsity issues or limiting context information (within the same sentence). Furthermore, the encoder and attention layers can be shared between features. By representing the encoder input as a combination of features (Alexandrescu and Kirchhoff, 2006), Sennrich and Haddow (2016) generalized the embedding layer in such a way that an arbitrary number of linguistic features can be explicitly integrated. They then investigated whether features such as lemmas, subword tags, morphological features, POS tags and dependency labels could be useful for NMT systems or whether their inclusion is redundant.

Similarly, on the syntax level, Shi et al. (2016) show that although NMT systems are able to partially learn syntactic information, more complex patterns remain problematic. Furthermore, sometimes information is present in the encoding vectors but is lost during the decoding phase (Vanmassenhove et al., 2017). Sennrich and Haddow (2016) show that the inclusion of linguistic features leads to improvements over the NMT baseline for EN–DE (0.6 BLEU), DE–EN (1.5 BLEU) and EN–RO (1.0 BLEU). When evaluating the gains from the features individually, it results that the gain from different features is not fully cumulative. Nadejde et al. (2017) extend their work by including CCG supertags as explicit features in a factored NMT systems. Moreover, they experiment with serializing and multitasking and show that tightly coupling the words with their syntactic features leads to improvements in translation quality (measured by BLEU) while a multitask approach (where the NMT predicts CCG supertags and words independently) does not perform better than the baseline system. A similar observation was made by Li et al (2017), who incorporate the linearized parse trees of the source sentences into ZH–EN NMT systems. They propose three different sorts of encoders: (a) a parallel RNN, (b) a hierarchical RNN, and (c) a mixed RNN. Like Nadejde et al. (2017), Li et al (2017) observe that the mixed RNN (the simplest RNN encoder), where words and label annotation vectors are simply stitched together in the input sequences, yields the best performance with a significant improvement (1.4 BLEU). Similarly, Eriguchi et al. (2016) integrated syntactic information in the form of linearized parse trees by using an encoder that computes vector representations for each phrase in the source tree. They focus on source-side syntactic information based on Head-Driven Phrase Structure Grammar (Sag et al., 1999) where target words are aligned not only with the corresponding source words but with the entire source phrase. Wu et al. (2017) focus on incorporating source-side long distance dependencies by enriching each source state with global dependency structure.

To the best of our knowledge, there has not been any work on explicitly integrating semantic information in NMT. Similarly to syntactic features, we hypothesize that semantic features in the form of semantic 'classes' can be beneficial for NMT providing it with an extra ability to generalize and thus better learn more complex semantico-syntactic patters.

3 Semantics and Syntax in NMT

3.1 Supersense Tags

The novelty of our work is the integration of explicit semantic features *supersenses* into an NMT system. Supersenses are a term which refers to the top-level hypernyms in the WordNet (Miller, 1995) taxonomy, sometimes also referred to as *semantic fields* (Schneider and Smith, 2015). The supersenses cover all nouns and verbs with a total of 41 supersense categories, 26 for nouns and 15 for verbs. To obtain the supersense tags we used the *AMALGrAM (A Machine Analyzer of Lexical Groupings and Meanings) 2.0* tool [1] which in addition to the noun and verb supersenses analyzes English input sentences for MWEs. An example of a sentence annotated with the AMALGrAM tool is given in (1):[2]

(1)

(a) "He seemed to have little faith in our democratic structures, suggesting that various articles could be misused by governments."

(b) "He **seemed|cognition** to have|stative little **faith|COGNITION** in our democratic structures|ARTIFACT , suggesting|communication that various articles|COMMUNICATION could be|'a misused|social by governments|GROUP ."

As can be noted in (1), some supersenses, such as *cognition* exist for both nouns and verbs. However, the supersense tags for verbs are always lowercased while the ones for nouns are capitalized. This way, the supersenses also provide syntactic information useful for disambiguation as in (2), where the word *work* is correctly tagged as a noun (with its capitalized supersense tag *ACT*) in the first part of the sentence and as a verb (with the lowercased supersense tag *social*). Furthermore, there is a separate tag to distinguish auxiliary verbs from main verbs.

(2)

(a) "In the course of my work on the opinion, I in fact became aware of quite a number of problems and difficulties for EU citizens who live and work in Switzerland"

(b) "In the course|EVENT of my work|ACT on the opinion|COGNITION , I **in_fact** became|stative aware of quite **a_number_of** problems|COGNITION and difficulties|COGNITION for **EU_citizens|GROUP** who live|social and work|social in Switzerland|LOCATION ."

Since MWEs and supersenses naturally complement each other, Schneider and Smith (2015) integrated the MWE identification task (Schneider et al., 2014) with the supersense tagging task of Ciaramita and Altun (2006). In Example (2), the MWEs *in fact*, *a number of* and *EU citizens* are retrieved by the tagger.

We add this semantico-syntactic information in the source as an extra feature in the embedding layer following the approach of Sennrich and Haddow (2016), who extended the model of Bahdanau et al. (2014). A separate embedding is learned for every source-side feature provided (the word itself, POS-tag, supersense tag etc.). These embedding vectors are then concatenated into one embedding vector and used in the model instead of the simple word embedding one (Sennrich and Haddow, 2016).

To reduce the number of out-of-vocabulary (OOV) words, we follow the approach of Sennrich et al. (2016) using a variant of BPE for word segmentation capable of encoding open vocabularies with a compact symbol vocabulary of variable-length subword units. For each word that is split into subword units, we copy the features of the word in question to its subword units. In (3), we give an example with the word 'stormtroopers' that is tagged with the supersense tag 'GROUP'. It is split into 5 subword units so the supersense tag feature is copied to all its five subword units. Furthermore, we add a *none* tag to all words that did not receive a supersense tag.

(3)

Input: "the stormtroopers"
SST: "the stormtroopers|GROUP"
BPE: "the stor@@ m@@ tro@@ op@@ ers"
Output: "the|**none** stor@@|**GROUP** ... op@@|**GROUP** ers|**GROUP**"

For the MWEs we decided to copy the supersense tag to all the words of the MWE (if provided by the tagger), as in (4). If the MWE did not receive a particular tag, we added the tag *mwe* to all its components, as in example (5)

(4)

Input: "EU citizens"
SST: "EU_ citizens|GROUP"
Output: "EU|GROUP citizens|GROUP"

(5)

Input: "a number of"
SST: "a_ number _ of"
Output: "a|**mwe** number|**mwe** of|**mwe** "

3.2 Supertags and POS-tags

We hypothesize that more general semantic information can be particularly useful for NMT in combination with more detailed syntactic information. To support our hypothesis we also experimented

[1] https://github.com/nschneid/pysupersensetagger

[2] All the examples are extracted from our data used later on to train the NMT systems

with syntactic features (separately and in combination with the semantic ones): POS tags and CCG supertags.

The POS tags are generated by the Stanford POS-tagger (Toutanova et al., 2003); for the supertags we used the EasySRL tool (Lewis et al., 2015) which annotates words with CCG tags. CCG tags provide global syntactic information on the lexical level. This kind of information can help resolve ambiguity in terms of prepositional attachment, among others. An example of a CCG-tagged sentence is given in (6):

(6)

It|NP is|(S[dcl]\NP)/NP a|NP/N modern|N/N form|N/PP of|PP/NP colonialism|N .|.

4 Experimental Set-Up

4.1 Data sets

Our NMT systems are trained on 1M parallel sentences of the Europarl corpus for EN–FR and EN–DE (Koehn, 2005). We test the systems on 5K sentences (different from the training data) extracted from Europarl and the newstest2013. Two different test sets are used in order to show how additional semantic and syntactic features can help the NMT system translate different types of test sets and thus evaluate the general effect of our improvement.

4.2 Description of the NMT system

We used the `nematus` toolkit (Sennrich et al., 2017) to train encoder-decoder NMT models with the following parameters: *vocabulary size:* 35000, *maximum sentence length:* 60, *vector dimension:* 1024, *word embedding layer:* 700, *learning optimizer:* `adadelta`. We keep the embedding layer fixed to 700 for all models in order to ensure that the improvements are not simply due to an increase of the parameters in the embedding layer. In order to by-pass the OOV problem and reduce the number of dictionary entries we use word-segmentation with BPE (Sennrich, 2015). We ran the BPE algorithm with 89, 500 operations. We trained all our BPE-ed NMT systems with CCG tag features, supersensetags (SST), POS tags and the combination of syntactic features (POS or CCG) with the semantic ones (SST). All systems are trained for 150,000 iterations and evaluated after every 10,000 iterations.

5 Results

5.1 English–French

For both test sets, the NMT system with supersenses (SST) converges faster than the baseline (BPE) NMT system. As we hypothesized, the benefits of the features added, was more clear on the newstest2013 than on the Europarl test set. Figure 1 compares the BPE-ed baseline system (BPE) with the supertag-supersensetag system (CCG–SST) automatically evaluated on the newstest2013 (in terms of BLEU (Papineni et al., 2002)) over all 150,000 iterations. From the graph, it can also be observed that the system has a more robust, consistent learning curve.

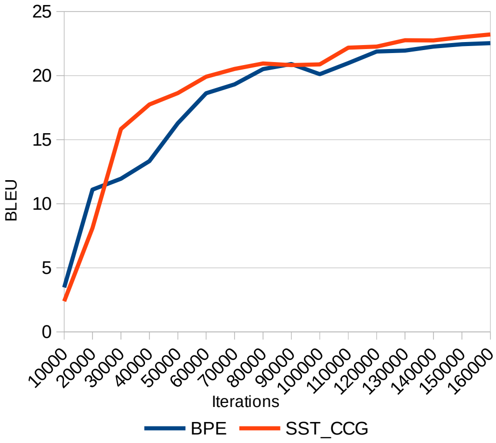

Figure 1: Baseline (BPE) vs Combined (SST–CCG) NMT Systems for EN–FR, evaluated on the newstest2013.

To see in more detail how our semantically enriched SST system compares to an NMT system with syntactic CCG supertags and how a system that integrates both semantic features and syntactic features (SST–CCG) performs, a more detailed graph is provided in Figure 2 where we zoom in on later stages of the learning process. Although Sennrich and Haddow (2016) observe that features are not necessarily cumulative (possibly since the information from the syntactic features partially overlapped), the system enriched with both semantic and syntactic features outperforms the two separate systems as well as the baseline system. The best CCG-SST model (23.21 BLEU) outperforms the best BPE-ed baseline model (22.54 BLEU) with 0.67 BLEU (see Table 1). Moreover, the benefit of syntactic and semantic features seems to be more than cumulative at some points, confirming the idea that providing both information sources can help the NMT system learn semantico-syntactic patterns. This supports our hypothesis that semantic and syntactic features are particularly useful when combined.

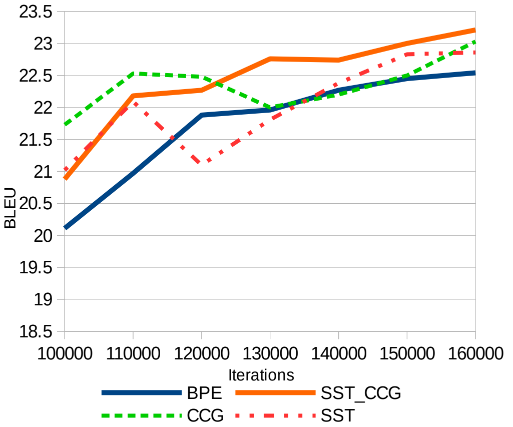

Figure 2: Baseline (BPE) vs Syntactic (CCG) vs Semantic (SST) and Combined (SST–CCG) NMT Systems for EN–FR, evaluated on the newstest2013.

BLEU	BPE	CCG	SST	SST–CCG
Best Model	22.54	23.03	22.86	23.21

Table 1: Best BLEU scores for Baseline (BPE), Syntactic (CCG), Semantic (SST) and Combined (SST–CCG) NMT systems for EN-FR evaluated on the newstest2013

5.2 English–German

The results for the EN–DE system are very similar to the EN–FR system: the model converges faster and we observe the same trends with respect to the BLEU scores of the different systems. Figure 3 compares the BPE-ed baseline system (BPE) with the NMT system enriched with SST and CCG tags (SST–CCG). In the last iterations, see Figure 4, we see how the two systems enriched with supersense tags and CCG tags lead to small improvements over the baseline. However, their combination (SST–CCG) leads to a more robust NMT system with a higher BLEU (see Table 2).

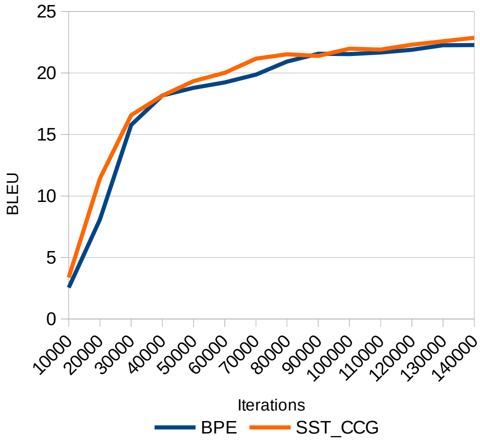

Figure 3: Baseline (BPE) vs Combined (CCG–SST) NMT Systems for English–German, evaluated on the Europarl test set.

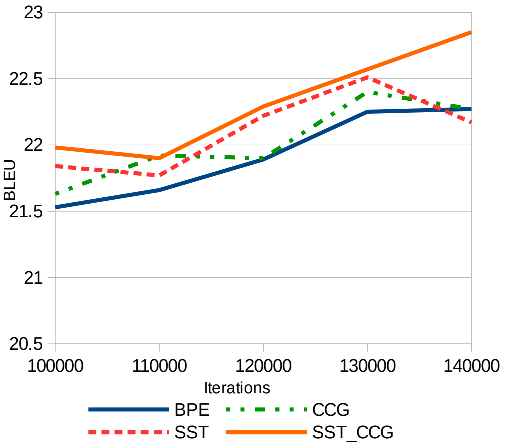

Figure 4: Baseline (BPE) vs Syntactic (CCG) vs Semantic (SST) and Combined (CCG–SST) NMT Systems for EN–DE, evaluated on the Europarl test set.

BLEU	BPE	CCG	SST	SST–CCG
Best Model	22.32	22.47	22.51	**22.85**

Table 2: Best BLEU scores for Baseline (BPE), Syntactic (CCG), Semantic (SST) and Combined (SST–CCG) NMT systems for EN-DE evaluated on the Europarl test set.

6 Conclusions and Future Work

In this work we experimented with EN–FR and EN–DE data augmented with semantic and syntactic features. For both language pairs we observe that adding extra semantic and/or syntactic features leads to faster convergence. Furthermore, the benefit of the additional features is more clear on a dissimilar test set which is in accordance with our original hypothesis stating that semantic and syntactic features (and their combination) can be beneficial for generalization. In the future, we would like to perform manual evaluations on the outputs of our systems to see whether they correlate with the BLEU scores. In the next step, we will let the models converge, create the ensemble models for the different systems and compute whether the increase in BLEU score is significant. Furthermore, we would like to experiment with larger datasets to verify whether the positive effect of the linguistic features remains.

7 Acknowledgements

I would like to thank Natalie Schluter for her insightful comments and feedback and Roam Analytics for sponsoring the ACL SRW travel grant.

This work has been supported by the Dublin City University Faculty of Engineering & Computing under the Daniel O'Hare Research Scholarship scheme and by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Roee Aharoni and Yoav Goldberg. 2017. Towards string-to-tree neural machine translation. In *Proceedings of the Association for Computational Linguistics, ACL*, Vancouver, Canada.

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored Neural Language Models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4, New York, USA.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of The Association for Computer Linguistics (ACL-08)*, pages 763–770, Ohio, USA.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. Banff, Canada.

Rafael E Banchs and Marta R Costa-Jussà. 2011. A Semantic Feature for Statistical Machine Translation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 126–134, Portland, Oregon, USA.

Joost Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph Convolutional Encoders for Syntax-Aware Neural Machine Translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 257–267, Austin, Texas, USA.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of EMNLP 2014*, pages 1724–1734, Doha, Qatar.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-Coverage Sense Disambiguation and Information Extraction with a Supersense Sequence Tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia.

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2016. Tree-to-Sequence Attentional Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 823–833, Berlin, Germany.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In *Proceedings of the Association for Computational Linguistics, ACL*, Vancouver, Canada.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word-Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, pages 664–674, Jeju Island, Republic of Korea.

Rejwanul Haque, Sudip Kumar Naskar, Yanjun Ma, and Andy Way. 2009. Using supertags as source language context in smt. In *European Association for Machine Translation*, Barcelona, Spain.

Hany Hassan, Khalil Sima'an, and Andy Way. 2007. Supertagged phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 288–295, Prague, Czech Republic.

Zhaopeng Tu Muhua Zhu Min Zhang Guodong Zhou Junhui Li, Deyi Xiong. 2017. Modeling Source Syntax for Neural Machine Translation. In *Proceedings of the Association for Computational Linguistics, ACL*, page 688697, Vancouver, Canada.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL, Volume 1: Long Papers*, pages 655–665, Baltimore, MD, USA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, Phuket, Thailand.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning Joint Meeting following ACL (EMNLP-CONLL*, pages 868–876, Prague, Czech Republic.

Mike Lewis, Luheng He, and Luke Zettlemoyer. 2015. Joint A* CCG Parsing and Semantic Role Labelling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1444–1454, Lisbon, Portugal.

David Mareček, Rudolf Rosa, Petra Galušččáková, and Ondřej Bojar. 2011. Two-step Translation with Grammatical Post-Processing. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 426–432, Edinburgh, UK.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *hlt-Naacl*, volume 13, pages 746–751, Atlanta, USA.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting Target Language CCG Supertags Improves Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Copenhagen, Denmark.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Ivan A Sag, Thomas Wasow, Emily M Bender, and Ivan A Sag. 1999. *Syntactic theory: A formal introduction*, volume 92. Center for the Study of Language and Information Stanford, CA.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut. volume 2, pages 193–206, Baltimore, Maryland, USA.

Nathan Schneider and Noah A. Smith. 2015. A Corpus and Model Integrating Multiword Expressions and Supersenses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*, pages 1537–1547, Denver, Colorado, USA.

Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics*, 3:169–182.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel L"aubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation, WMT, ACL*, pages 83–91, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL Volume 1: Long Papers*, Berlin, Germany.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of Empirical Methods on Natural Language Processing (EMNLP 2016)*, Austin, Texas, USA.

Mark Steedman. 2000. *The Syntactic Process*, volume 24. Cambridge: MIT Press.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pages 3104–3112, Montreal, Quebec, Canada.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *ACL*, pages 514–522.

Eva Vanmassenhove, Jinhua Du, and Andy Way. 2017. Investigating 'aspect' in nmt and smt: Translating the english simple past and present perfect. *Computational Linguistics in the Netherlands Journal*, 7:109–128.

Philip Williams and Philipp Koehn. 2012. Ghkm Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 388–394.

Shuangzhi Wu, Ming Zhou, and Dongdong Zhang. 2017. Improved Neural Machine Translation with Source Syntax. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017).*, pages 4179–4185, Melbourne, Australia.