Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings

Annette Rios[⋆] and **Laura Mascarell**[⋆] and **Rico Sennrich**[†⋆]
[⋆]Institute of Computational Linguistics, University of Zurich
[†]School of Informatics, University of Edinburgh

Abstract

Word sense disambiguation is necessary in translation because different word senses often have different translations. Neural machine translation models learn different senses of words as part of an end-to-end translation task, and their capability to perform word sense disambiguation has so far not been quantified. We exploit the fact that neural translation models can score arbitrary translations to design a novel cross-lingual word sense disambiguation task that is tailored towards evaluating neural machine translation models. We present a test set of 7,200 lexical ambiguities for German→English, and 6,700 for German→French, and report baseline results. With 70% of lexical ambiguities correctly disambiguated, we find that word sense disambiguation remains a challenging problem for neural machine translation, especially for rare word senses. To improve word sense disambiguation in neural machine translation, we experiment with two methods to integrate sense embeddings. In a first approach we pass sense embeddings as additional input to the neural machine translation system. For the second experiment, we extract lexical chains based on sense embeddings from the document and integrate this information into the NMT model. While a baseline NMT system disambiguates frequent word senses quite reliably, the annotation with both sense labels and lexical chains improves the neural models' performance on rare word senses.

1 Introduction

Semantically ambiguous words present a special challenge to machine translation systems: in order to produce a correct sentence in the target language, the system has to decide which meaning is accurate in the given context. Errors in lexical choice can lead to wrong or even incomprehensible translations. However, quantitatively assessing errors of this type is challenging, since automatic metrics such as BLEU (Papineni et al., 2002) do not provide a sufficiently detailed analysis.

Several ways of evaluating lexical choice for machine translation have been proposed in previous work. Cross-lingual lexical choice tasks have been created for the evaluation of word sense disambiguation (WSD) systems (Mihalcea et al., 2010; Lefever and Hoste, 2013), and have been applied to the evaluation of MT systems (Carpuat, 2013). Vickrey et al. (2005) evaluate lexical choice in a blank-filling task, where the translation of an ambiguous source word is blanked from the reference translation, and an MT system is tested as to whether it can predict it. In all these tasks, a word-level translation (or set of translations) is defined as the gold label. A major problem is that an MT system will be punished for producing a synonym, paraphrase, or inflected variant of the predefined gold label. We thus propose a more constrained task where an MT system has to select one out of a predefined set of translations.

Neural machine translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2015) has recently emerged as the new state of the art in machine translation, producing top-ranked systems in recent shared tasks (Luong and Manning, 2015; Sennrich et al., 2016a; Neubig, 2016). The strengths and weaknesses of NMT have been the subject of recent research, and previous studies involving human analysis have consistently found NMT to be

Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 11–19
Copenhagen, Denmark, September 711, 2017.

more fluent than phrase-based SMT (Neubig et al., 2015; Bojar et al., 2016; Bentivogli et al., 2016), but results in terms of adequacy are more mixed. Bentivogli et al. (2016) report improvements in lexical choice based on HTER matches on the lemma-level, while Bojar et al. (2016) found no clear improvement in a direct assessment of adequacy. Neubig et al. (2015) perform an error annotation in which the number of lexical choice errors even increases slightly by reranking a syntax-based statistical machine translation system with an NMT model.

We aim to allow for a large-scale, reproducible method of assessing the capability of an NMT model to perform lexical disambiguation. NMT systems can not only be used to generate translations of a source sentence, but also to assign a probability $P(T|S)$ for any given pair of a source sentence S and a target sentence T. We use this feature to create a test set with artificially introduced lexical disambiguation errors. Comparing the scores of an NMT model on these contrastive translations to the score of the reference allows us to assess how well the model can distinguish different senses in ambiguous words.

We have created two test sets for the language pairs German-English and German-French with about 6,500 and 6,700 sentence pairs respectively.[1] Based on the performance of state-of-the-art NMT systems on these test sets, we discuss the capability of NMT to perform lexical disambiguation.

Furthermore, we present two methods to improve word sense disambiguation in neural machine translation by allowing the model to learn sense-specific word embeddings. Both methods are based on an external word sense disambiguation. While the first method passes sense labels as additional input to an NMT system, the second is motivated by the hypothesis that document-level context is valuable for disambiguation. We model this context via lexical chains, i.e. sequences of semantically-similar words in a given text that express the topic of the segment they cover in a condensed form. Our method is inspired by Galley and McKeown (2003), who present an approach to build English lexical chains automatically using WordNet (Miller, 1995) and evaluate its performance on a sense disambiguation task. Instead of WordNet, we use sense embeddings in order to determine the similarity between the words in a document and thus find and annotate the lexical chains. Experimental results show the potential of lexical chains at disambiguating word senses.

[1]The test set is available from https://github.com/a-rios/ContraWSD.

2 Contrastive Translations

The test set consists of sentence pairs that contain at least one ambiguous German word. In order to produce contrastive translation pairs, we create an automatically modified version of the reference translation where we replace the original translation of a given ambiguous word with the translation of one of its other meanings. We cluster different translations that overlap in meaning, i.e. that are (at least sometimes) used interchangeably. We do not produce any contrastive translations that belong to the same cluster as the reference translation.

As an example, we show the sense clusters that we consider for two ambiguous German words:

Schlange:	
serpent, snake	
line, queue	
Abzug:	
withdrawal, departure	*rétraction, sortie*
trigger	*gâchette*
discount, subtraction	*déduction, soustraction*

Table 1 shows an example of source, reference, and contrastive sentences.

Our approach is inspired by Sennrich (2017), who use contrastive translation pairs to evaluate various error types, including morpho-syntactic agreement and polarity errors. Apart from focusing on another error type, namely word sense errors, our approach differs in that we pair a human reference translation not just with one contrastive example, but a set of contrastive examples, i.e. a set of incorrect translations of the semantically ambiguous source word. The model is considered correct if it scores the human reference translation higher than all of the contrastive translations. Note that this evaluation does not directly assess the translation output of a system, which might be different from the set of translations that are scored, or the search performance of a system. Instead, its focus is to identify specific model errors.

3 Lexical Choice Errors

In a first step, we compile a list of German nouns that have semantically distinct translations in English and French from the lexical translation tables

of existing German-English and German-French phrase-based MT systems, and we clean these lists manually. We then extract sentence pairs from parallel corpora for all ambiguous words in our lists. Since for most ambiguous words, one or more of their meanings are relatively rare, a large amount of parallel text is necessary to extract a sufficiently balanced number of examples.[2]

When creating the test set, our goal is to produce contrastive translations that cannot be easily identified as wrong based on grammatical or phonological features. We do not consider ambiguities across word classes (*Flucht* - 'flight, escape' vs. *flucht* - 'he/she curses'). Furthermore, we do not consider German words with different meanings distinguished by gender (*der Leiter* (m.) - 'leader' vs. *die Leiter* (f.) - 'ladder').

Contrastive translations are produced automatically based on a replacement of the target word with the specified contrastive variants. We ensure that contrastive translations match the original translation in number; in French, we also limit replacements to those that match the original translation in gender, and take into account elision for vowel-initial words.

We consider both plural and singular forms in German, but exclude word forms that are unambiguous. For instance, the German singular word *Schuld* can refer to *debt* or *guilt*, however, the plural form *Schulden* can only be translated as *debts*.

Furthermore, we exclude a small number of cases where the context in either source or target sentence clearly indicates the meaning: For instance, if the German word *Absatz* ('heel', 'sales', 'paragraph') is followed by a number, the translation is in all likelihood 'paragraph' and contrastive sentences with 'heel' or 'sales' will not present a challenge for the model.

Following our strategy of focusing on difficult cases, we oversample the less frequent word senses for the test set to reduce the performance of a simple most frequent sense baseline to that of random guessing. Specifically, we include 100 test instances per word sense, or the total amount of available sentence pairs if less than 100 were found in the parallel data.

For German-English, the test set contains 84 word senses, with on average 3.5 contrastive translations per reference; for German-French, it contains 71 word senses, with an average of 2.2 contrastive translations per reference. A full list of word senses can be found in the appendix.

We include the location of the sentence in the original corpus in our metadata to allow future experiments with document-level information.[3]

4 Sense Embeddings in Neural Machine Translation

In addition to the evaluation of a standard NMT model on the word sense disambiguation task detailed in the previous section, we present two experiments on German→English and German→French to improve lexical choice using methods from WSD. In a first approach, we compute sense embeddings and include the resulting sense labels into the NMT model as additional input features (Alexandrescu and Kirchhoff, 2006; Sennrich and Haddow, 2016). For our second experiment, instead of adding the labels directly to the input, we use them to build lexical chains of similar words in the given document. These lexical chains contain information about the topic and/or domain of the document, and we include them as additional features into our NMT model.

4.1 Sense Embeddings

Sense embeddings are vector representations of word senses in a vector space, but unlike word embeddings, where every word form receives a vector representation, with sense embeddings we obtain separate vector representations for each sense of a given word. To compute the sense embeddings we

[2] Sentence pairs have been extracted from the following corpora:

- WMT test and development sets 2006-2016 (de-en) and 2006-2013 (de-fr)
- Crédit Suisse News Corpus https://pub.cl.uzh.ch/projects/b4c/de/
- corpora from OPUS ((Tiedemann, 2012)):
  - Global Voices (http://opus.lingfil.uu.se/GlobalVoices.php)
  - Books (http://opus.lingfil.uu.se/Books.php)
  - EU Bookshop Corpus (http://opus.lingfil.uu.se/EUbookshop.php)
  - OpenSubtitles 2016 (German-French) (http://opus.lingfil.uu.se/OpenSubtitles2016.php)
- MultiUN (Ziemski et al., 2016)

[3] A snapshot of the corpora used to extract the examples can be found at http://data.statmt.org/ContraWSD/.

source:	*Also nahm ich meinen amerikanischen Reisepass und stellte mich in die **Schlange** für Extranjeros.*
reference:	*So I took my U.S. passport and got in the **line** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **snake** for Extranjeros.*
contrastive:	*So I took my U.S. passport and got in the **serpent** for Extranjeros.*
source:	*Er hat zwar schnell den Finger am **Abzug**, aber er ist eben neu.*
reference:	*Il a la **gâchette** facile mais c'est parce qu'il débute.*
contrastive:	*Il a la **soustraction** facile mais c'est parce qu'il débute.*
contrastive:	*Il a la **déduction** facile mais c'est parce qu'il débute.*
contrastive:	*Il a la **sortie** facile mais c'est parce qu'il débute.*
contrastive:	*Il a la **rétraction** facile mais c'est parce qu'il débute.*

Table 1: Contrastive Translations

use $SenseGram$[4] (Pelevina et al., 2016), which has been shown to perform as good as stat-of-the-art unsupervised WSD systems.

The method to learn the sense embeddings using $SenseGram$ consists of four steps that we briefly summarise here. First, the method learns word embeddings using the $word2vec$ toolkit (Mikolov et al., 2013).[5] It then uses these word embeddings to build a word similarity graph, where each word is linked to its 200 nearest neighbours. Next, it induces a sense inventory, where each sense is represented by a cluster of words (e.g. the sense of *table-furniture* is represented with the word cluster *desk*, *bench*, *dining table*, *surface*, and *board*). The sense inventory of each word is obtained through clustering the ego-networks of its related words. Finally, the method computes the sense embedding of each word sense by averaging the vectors of the words in the corresponding cluster.

Once the sense embeddings are learned, we label all content words in the data with their corresponding sense and include this information as additional features.

4.2 Lexical Chains

As described above, $SenseGram$ allows us to disambiguate a word based on the context in which it occurs. Based on the disambiguated words, we can detect the lexical chains, i.e. chains of semantically similar words within a given document. To compute the semantic similarity between two word senses, we calculate the cosine similarity between their sense embeddings.[6] The closer to 1.0 the resulting value is, the higher their semantic similarity. To distinguish between similar and non-similar senses, we set a threshold of 0.85 that we manually picked by looking at how different values affect the resulting lexical chains: a lower threshold builds lexical chains containing sense words that are not sufficiently related, whereas a higher threshold results in semantically strong, but possibly incomplete lexical chains that do not cover all words belonging to the chain.

We use the method proposed by Mascarell (2017) to detect lexical chains in a document. This method is inspired by Morris and Hirst (1991)'s approach, which manually finds lexical chains in a document using a thesaurus to obtain the similarity between words. As detailed in Section 4.1, we use sense embeddings instead of a dictionary to compute the semantic similarity.

Given a document as input, our method processes sentences and their content words sequentially. For each sentence, it computes the semantic similarity between the current content word *c* and each previous content word *c'* in the previous five sentences, based on the approach by Morris and Hirst (1991). If *c* and *c'* are semantically similar, our method proceeds as follows:

- If *c* and *c'* are not part of a chain, create a new chain with *c* and *c'*.
- If *c'* is in a chain ch_i, append *c* to ch_i.

[4] https://github.com/tudarmstadt-lt/sensegram

[5] Embeddings for our models were learned on the following corpora:

- SdeWaC (Faaß and Eckart, 2013) (∼768M words)
- Common Crawls (∼775M words)
- Europarl (∼47M words)
- News Commentary (∼6M words)

[6] Using sense embeddings instead of word embeddings for this task ensures that we can recognize similar words even if they are polysemic and not all of their senses are related. For instance, *mouse* and *rat* are related if *mouse* refers to the animal, but not if *mouse* refers to the computer device.

- If *c* and *c'* are in two different chains, merge both chains.

Since every linked word in the chain provides context for disambiguation, the method creates as many links as possible between similar words. Therefore, it also preserves one-transitive links: c_i links to c_{i+l} by transitivity if c_i links to c_{i+k} and c_{i+k} to c_{i+l}, where $i<k<l$ (Morris and Hirst, 1991).

As Morris and Hirst (1991) indicate, words linked by one-transitive links are semantically related, but words further apart in the chain might not be: In their paper, they point to the lexical chain {*cow*, *sheep*, *wool*, *scarf*, *boots*, *hat*, *snow*}. While consecutive words in the chain such as *wool* and *scarf* are semantically related, *cow* and *snow* are not.

To provide the NMT model with the detected lexical chains in the source, we represent this discourse knowledge in the input as a combination of features. Accordingly, each word in the lexical chain is annotated with its linked words as factors. For example, if the German word *Absatz* is linked in the lexical chain to *Wirtschaft* ('economy') and *Verkauf* ('sale'), it is represented as *Absatz*|*Wirtschaft*|*Verkauf*. The resulting vector representation of *Absatz* is the vector concatenation of each individual feature's embeddings.

Since all words in the input must have the same number of factors, each word that is not part of a lexical chain is annotated with itself as factors. Similarly, words linked to only one word are annotated with the corresponding linked word in the chain and the word itself.

5 Evaluation

We present an evaluation with two basic neural MT systems, trained with Nematus (Sennrich et al., 2017), using byte pair encoding (BPE) on both source and target side (Sennrich et al., 2016b). For both the German-English and the German-French experiments, we train a model on 2.1 million sentence pairs from Europarl (v7) and News Commentary (v11).[7] We use these corpora because they contain document boundaries, which is a requirement for the lexical chains experiments.

We present further results for models that use additional source-side features, a) the sense labels themselves and b) lexical chains. The feature is given its own embedding space, and the model can thus learn sense-specific embeddings. If a word is segmented into multiple subword units by BPE, the additional input feature of the word is repeated for each unit. Vocabulary size for all models is 90,000.

We train the models for a week, using Adam (Kingma and Ba, 2015) to update the model parameters on minibatches of the size 80. Every 10,000 minibatches, we validate our model on a held out development set via BLEU and perplexity. The maximum length of the sentences is 50. The total size of the embedding layer is 500 for both the baseline and the system trained with additional input features, and the dimension of the hidden layer is 1024. For the experiments with additional input features, we divide the embedding size equally among the features. Conceivably, keeping the dimensionality of the word embedding constant and adding more parameters for additional features would result in better performance, but we wanted to rule out that any performance improvements are solely due to an increase in model size.

To assess a model's capability to distinguish different meanings of ambiguous words, we let it assign a score to the reference translation and to the artificially created contrastive translations. If the score of the reference translation is higher than the scores of all contrastive translations, this counts as a correct decision.

system	accuracy
de-en ($N = 7243$)	
NMT baseline	0.7095
NMT sense labels	0.7138
NMT lexical chains	0.7034
human	≈0.96
de-fr ($N = 6746$)	
NMT baseline	0.7023
NMT sense labels	0.6998
NMT lexical chains	0.7083
human	≈0.93

Table 2: Word sense disambiguation accuracy

German	*Sehen Sie die **Muster**?*
reference	*Do you see the **patterns**?*
contrastive	*Do you see the **examples**?*

Table 4: Ambiguous sentence pair

[7] http://opus.lingfil.uu.se/News-Commentary11.php

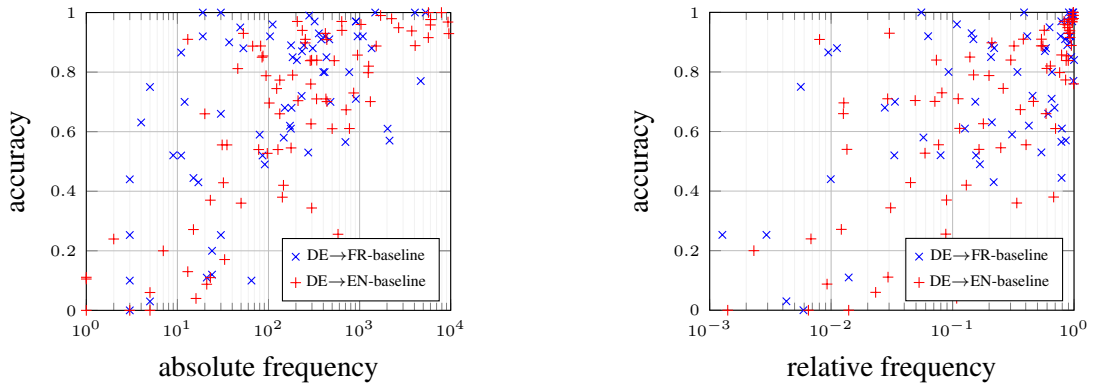

Figure 1: Word sense disambiguation accuracy by word sense frequency in training set (absolute, or relative to source word frequency).

		de-en				de-fr		
		baseline	sense labels	lexical chains		baseline	sense labels	lexical chains
frequency	senses*	accuracy	accuracy	accuracy	senses*	accuracy	accuracy	accuracy
>10000	2	0.9840	0.9840	0.9840	2	0.9900	0.9900	0.9900
>5000	7	**0.9639**	0.9534	0.9459	1	1.0000	1.0000	0.9900
>2000	4	**0.9386**	0.9284	0.9284	3	0.7375	**0.7725**	0.7150
>1000	6	0.8598	**0.8632**	0.8427	3	0.9333	**0.9367**	0.9167
>500	8	**0.7410**	0.7308	0.7090	6	0.8260	0.8260	**0.8361**
>200	17	0.7800	0.7734	**0.7900**	16	0.8444	**0.8475**	0.8406
>100	9	0.6058	0.6095	**0.6156**	9	**0.7544**	0.7456	0.6933
>50	8	**0.7899**	0.7645	0.7630	6	0.5160	0.5200	**0.6420**
>20	9	0.4055	**0.4521**	0.3945	8	0.5276	0.5430	**0.5469**
0-20	14	0.3127	**0.3664**	0.3237	17	0.4924	0.4611	**0.5156**

Table 3: Accuracy of word sense prediction by frequency of word sense in training set (* number of senses in frequency range).

	baseline	sense labels	lexical chains
de-en	17.1	16.9	17.1
de-fr	14.6	14.6	14.7

Table 5: Average BLEU scores on newstest 2009-2013

As Table 2 shows, both the German→French and the German→English baseline model achieve an accuracy of 0.70 on the test set. We also report accuracy of a smaller-scale human evaluation, in which two human annotators (one per language pair) were asked to identify the correct translation for a random sample of the test set (N=100–150). The annotation was performed purely on sentence-level, without any document context, and shows that some ambiguities are even hard for a human to resolve without context. Consider the sentence pairs in Table 4 for such an example. We speculate that both humans and MT systems should be able to resolve more ambiguities with wider context. Even with only sentence-level information, the gap between human and NMT performance is sizeable, between 23 and 26 percentage points.

An important indicator of how well a word sense is translated by NMT is its frequency in the training data. Figure 1 illustrates the relationship between the frequency of a word sense in the training data (both absolute and relative to the frequency of the source word) and the accuracy the model achieves on the test set.

There is a high correlation between word sense frequency and accuracy: for German→English, Spearman's ρ is 0.75 for the correlation between accuracy and absolute frequency, and 0.77 for the correlation between accuracy and relative frequency. For German→French, ρ is 0.58 for both. It is unsurprising that the most frequent word sense is preferred by the model, and that accuracy for it is high. We hence want to highlight performance on rarer word senses. Table 3 shows the word sense accuracy of the NMT models grouped by frequency classes and the number of senses in each

class. All models achieve close to 100% accuracy on words that occur more than 10,000 times in the training data. For the rare senses however, the NMT models are much less reliable: for word senses seen 0-20 times in training, the baseline accuracy is between 31-49%.

The annotation of the source side with sense labels improves the accuracy on the test set by 0.43% for German→English, while the lexical chains does not improve the model on average. On the other hand for German→French, the lexical chains result in an improvement of 0.6%, but the annotation with sense labels does not lead to a better score on the test set on average. As shown in Table 3, there is little room for improvement for frequent word senses, and sense labels and lexical chains show the strongest improvements over the baseline for the less frequent word senses. Table 5 contains the average BLEU scores on the newstest 2009-2013 test sets.

6 Conclusions

This paper introduces a novel lexical decision task for the evaluation of NMT models, and presents test sets for German-English and German-French. This task allows for the automatic and quantitative analysis of the ability of NMT models to perform lexical disambiguation, a phenomenon that has previously been remarked to be challenging for NMT. First evaluations with NMT models show that lexical choice is resolved well for frequent word senses, but not for infrequent word senses. Additional experiments to add a) sense labels to content words and b) topic knowledge in the form of lexical chains to the NMT model shows that semantic information improves lexical choice especially for word senses that do not occur frequently in the training data. We find that the inclusion of sense labels improves lexical choice on our test set 0.43% for German→English. Furthermore, we gain a small increase of 0.6% in accuracy with lexical chains for German→French.

We consider the performance of the baseline NMT systems respectable, given that the test set was created to be challenging, and has a strong focus on difficult cases. Our experiments indicate that NMT models perform poorly for rare word senses, and we observe moderate improvements for these rare word senses by using methods from WSD to complement the disambiguation capability of the main NMT model. Still, the problem is far from solved, and there is a sizeable difference of 23-26 percentage points between NMT performance and human performance. We hope that the release of our test set will inspire and support future research on the problem of word sense disambiguation for machine translation. In our human experiments, we also found evidence that wider document context is necessary to solve this task.

Acknowledgments

We are grateful to the Swiss National Science Foundation (SNF) for support of the project CoNTra (grant number 105212_169888) and the Synergia MODERN project (grant number 147653).

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)* . San Diego, CA, USA.

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*. Berlin, Germany, pages 131–198.

Marine Carpuat. 2013. A Semantic Evaluation of Machine Translation Lexical Choice. In *Proceedings of the 7th Workshop on Syntax, Semantics and Structure in Statistical Translation*. Atlanta, Georgia, pages 1–10.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – A Corpus of Parsable Sentences from the Web. In *Language Processing and Knowledge in the Web*, Springer Berlin Heidelberg, volume 8105, pages 61–68.

Michel Galley and Kathleen McKeown. 2003. Improving Word Sense Disambiguation in Lexical Chaining. In *Proceedings of the 18^{th} International Joint Conference on Artificial Intelligence (IJCAI-03)*. Acapulco, Mexico, pages 1486–1488.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: a Method for Stochastic Optimization. In *Proceedings of the 3^{rd} International Conference on Learning Representations ((ICLR2015))*. San Diego, CA, USA.

Els Lefever and Véronique Hoste. 2013. SemEval-2013 Task 10: Cross-lingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA, pages 158–166.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation 2015*. Da Nang, Vietnam.

Laura Mascarell. 2017. Lexical Chains meet Word Embeddings in Document-level Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*. Copenhagen, Denmark.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 Task 2: Cross-lingual Lexical Substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Los Angeles, California, pages 9–14.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations (ICLR 2013)*. Scottsdale, AZ, USA.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

Jane Morris and Graeme Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17(1):21–48.

Graham Neubig. 2016. Lexicons and Minimum Risk Training for Neural Machine Translation: NAIST-CMU at WAT2016. In *Proceedings of 3rd Workshop on Asian Translation (WAT 2016)*. Osaka, Japan.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural Reranking Improves Subjective Quality of Machine Translation: NAIST at WAT2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*. Kyoto, Japan, pages 35–41.

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*. Philadelphia, Pennsylvania, pages 311–318.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making Sense of Word Embeddings. In *Proceedings of the 1^{st} Workshop on Representation Learning for NLP*. Berlin, Germany, pages 174–183.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain.

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

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, volume 1 Research Papers, pages 83–91.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*. Berlin, Germany, volume 2 Shared Task Papers, pages 371–376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, pages 1715–1725.

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 2014*. Montreal, Quebec, Canada, pages 3104–3112.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 2214–2218.

David Vickrey, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-Sense Disambiguation for Machine Translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Vancouver, British Columbia, Canada, pages 771–778.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 3530–3534.