# The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation

**Alessandro Raganato**[*†], **Yves Scherrer**[*] **and Jörg Tiedemann**[*]

[*]University of Helsinki [†]Basement AI
{name.surname}@helsinki.fi

## Abstract

Supervised Neural Machine Translation (NMT) systems currently achieve impressive translation quality for many language pairs. One of the key features of a correct translation is the ability to perform word sense disambiguation (WSD), i.e., to translate an ambiguous word with its correct sense. Existing evaluation benchmarks on WSD capabilities of translation systems rely heavily on manual work and cover only few language pairs and few word types. We present MuCoW, a multilingual contrastive test suite that covers 16 language pairs with more than 200 000 contrastive sentence pairs, automatically built from word-aligned parallel corpora and the wide-coverage multilingual sense inventory of BabelNet. We evaluate the quality of the ambiguity lexicons and of the resulting test suite on all submissions from 9 language pairs presented in the WMT19 news shared translation task, plus on other 5 language pairs using pretrained NMT models. The MuCoW test suite is available at http://github.com/Helsinki-NLP/MuCoW.

## 1 Introduction

Neural Machine Translation (NMT) has provided impressive advances in translation quality, leading to a discussion whether translations produced by professional human translators can still be distinguished from the output of NMT systems, and to what extent automatic evaluation measures can reliably account for these differences (Hassan Awadalla et al., 2018; Läubli et al., 2018; Toral et al., 2018). One answer to this question lies in the development of so-called *test suites* (Burchardt et al., 2017) or *challenge sets* (Isabelle et al., 2017) that focus on particular linguistic phenomena that are known to be difficult to evaluate with simple reference-based metrics such as BLEU. Existing test suites focus e.g. on morphosyntactic and syntactic divergences between source and target language (Burchardt et al., 2017; Burlot and Yvon, 2017; Isabelle et al., 2017; Sennrich, 2017; Burlot et al., 2018; Macketanz et al., 2018) or on discourse phenomena (Guillou and Hardmeier, 2016; Bawden et al., 2018; Müller et al., 2018; Guillou et al., 2018).

Another linguistic phenomenon that is challenging for translation is lexical ambiguity (Liu et al., 2018; Marvin and Koehn, 2018), i.e., words of the source language that have multiple translations in the target language representing different meanings. Recently, Rios Gonzales et al. (2017) introduced a lexical ambiguity benchmark called ContraWSD that is based on contrastive translation pairs: a sentence containing an ambiguous source word is paired with the correct reference translation and with a modified translation in which the ambiguous word has been replaced by a word of a different sense. Contrastive evaluation makes use of the ability of NMT systems to score given translations: a contrast is considered successfully detected if the reference translation obtains a higher score than an artificially modified translation.

However, all these test suites require significant amounts of expert knowledge and manual work for identifying the divergences and compiling the examples, which typically limits their coverage to a small number of language pairs and directions. For example, the test sets built by Rios Gonzales et al. (2017) cover only 65 ambiguous words for two language pair directions.

In this paper, we present a language-independent method for automatically building ContraWSD-style test suites. It involves the following steps: (1) identify ambiguous source words and their translations; (2) cluster the translations into senses; (3) select sentences with ambiguous words and create contrast pairs.

The setup proposed by Rios Gonzales et al.



*Proceedings of the Fourth Conference on Machine Translation (WMT)*, Volume 2: Shared Task Papers (Day 1) pages 470–480
Florence, Italy, August 1-2, 2019. 

| | | |
|---|---|---|
| 177 input | 26 documents | 9 system |
| 50 typing | 21 petition | 8 entered |
| 29 entering | 17 data | 8 command |
| 28 entry | 14 submission | 7 display |
| 27 loading | 13 the | 7 to |
| 26 enter | 11 inputting | ... |

Table 1: English words aligned with the German word *Eingabe* and their alignment frequencies. Words with frequency $< 10$ are discarded from further processing.

| | |
|---|---|
| Petition, Antrag, Gesuch, Eingabe | **petition**, request, postulation |
| Produktionsfaktor, Ressource, Eingabe | factors of production, **input**, resource |
| Eingabe (Computer), Dateneingabe, Input | **input**, data entry |

Table 2: Three bilingual German–English clusters for the German word *Eingabe*, as obtained from BabelNet. Intersected words with Table 1 are displayed in bold. The second and third clusters are merged because of the shared English word *input*.

(2017) has shown a certain number of drawbacks. First, it cannot be used in conjunction with online systems (which do not provide an API for scoring) or with rule-based systems. Second, it is unclear to what extent the score of an MT system reflects its quality, as it might never have generated that particular sentence. Third, it requires the explicit construction of contrastive sentences, which is not trivial, especially for morphologically rich languages. For these reasons, the WMT test suite calls focus on *translation test suites*, where the participants are asked to produce translations of the source sentence instead of scoring given hypotheses. Following Rios et al. (2018) and Macketanz et al. (2018), who proposed small-scale translation test suites targeting WSD, we participated at WMT with modified versions of MuCoW. The modifications only concern step (3).

As a result, we make available two variants of MuCoW, a **mu**ltilingual **co**ntrastive **w**ord sense disambiguation test suite for machine translation. The scoring variant covers 11 language pairs with a total of almost 240 000 sentence pairs. The translation variant covers 9 language pairs with a total of 15 600 sentences. The data and scoring scripts are available at https://github.com/Helsinki-NLP/MuCoW.

## 2 Building MuCoW

In this section, we describe the three steps needed to create a MuCoW test suite and illustrate them with some German→English examples.

### 2.1 Step 1: Identify ambiguous source words and their translations

We first compile a list of source language words that have a large number of distinct translations. For this, we apply the *eflomal* word alignment tool (Östling and Tiedemann, 2016) on a collection of parallel corpora, keeping only those source words that were aligned at least 10 times each with at least two distinct target words. We use parallel corpora from the OPUS collection (Tiedemann, 2012),[1] counting only one-to-one word alignment links. Table 1 provides an example.

### 2.2 Step 2a: Cluster target words via BabelNet

For each source word of the previous step, those target words that potentially share the same meaning (for example synonyms) are clustered together. To this end, we exploit BabelNet (Navigli and Ponzetto, 2012), a wide-coverage multilingual encyclopedic dictionary obtained automatically from various resources (WordNet and Wikipedia, among others). BabelNet 4.0 covers 284 languages with almost 16 million entries, called Babel synsets. Each entry represents a given meaning and includes a set of synonyms (synset) in different languages. Conveniently, it provides inter-resource mappings in multiple languages, which enables us to translate words and senses between several languages.

We query BabelNet with each source word and take the intersection of the alignment-inferred target words and the BabelNet-inferred target words. Crucially, we group the remaining target words according to the BabelNet sense clusters. Finally, we combine those clusters that share at least one common target word. Table 2 shows an example.

[1] We use the following corpora: Books v1, EU Bookshop Corpus v2, Europarl v7 (Koehn, 2005), MultiUN v1 (Eisele and Chen, 2010), News-Commentary v11, OpenSubtitles v2018 (Lison and Tiedemann, 2016), SETIMES v2 (Tyers and Alperen, 2010), Tatoeba v2, TED2013 v1.1 (Cettolo et al., 2013).

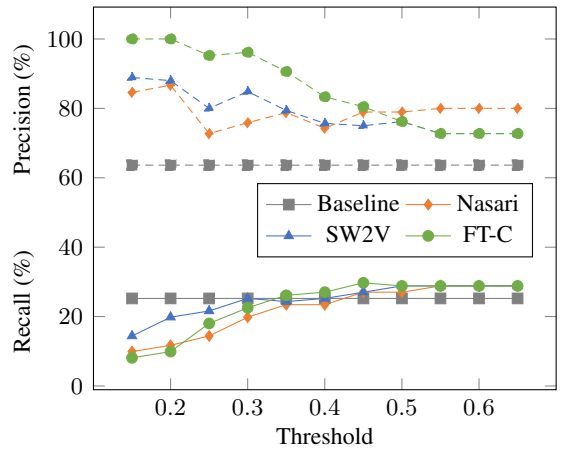


Figure 1: Precision (dashed) and recall (solid lines) values for different sense embeddings and thresholds.

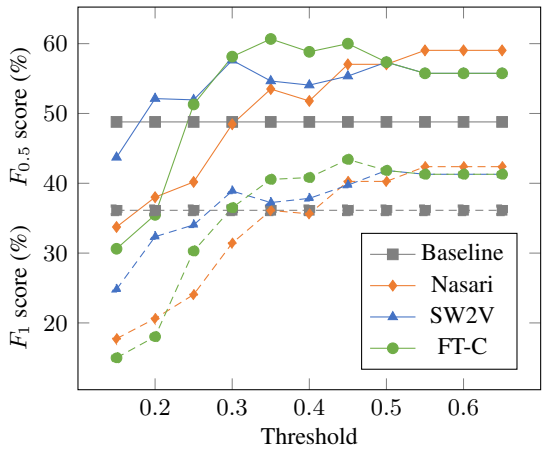


Figure 2: $F_1$-scores (dashed) and $F_{0.5}$-scores (solid lines) for different sense embeddings and thresholds.

### 2.3 Step 2b: Refine sense clusters with sense embeddings

It is known that lexical resources such as BabelNet tend to suffer from overly fine granularity of their sense inventory (Navigli, 2006; Palmer et al., 2007). We therefore introduce an additional merging step: i) we associate each Babel synset with an embedding, ii) compute pairwise cosine similarities between synsets, iii) and merge them if their embedding similarity is higher than a threshold $\gamma$.

Choosing a good Babel synset embedding and an optimal threshold is a difficult task. We evaluated three Babel synset vector representations, using the existing German→English ContraWSD test suite as gold standard:

**Nasari** (Camacho-Collados et al., 2016) is a vector representation built by combining the knowledge from Wikipedia and WordNet with word embeddings.

**SW2V** (Mancini et al., 2017) is a neural model that learns word and synset embeddings in a shared vector space exploiting a shallow graph-based disambiguation algorithm.

**FastText-Centroid (FT-C):** We also include a synset embedding representation by looking up the FastText word embeddings (Bojanowski et al., 2017) for all words in a synset and computing their centroid.

Note that Nasari and SW2V embeddings are tied to the (language-independent) BabelNet synset IDs and can therefore be applied in a straightforward way to non-English target languages.[2] As a baseline, we use the synset clusters obtained from Section 2.2.

We compute precision and recall scores for all three embedding methods with $\gamma$ threshold values ranging from 0.15 to 0.65 with a 0.05 step size. An inferred synset was considered correct if all its lexicalisations (if present) occurred in a single gold synset, and no lexicalisations of a gold synset were found in a different inferred synset. In other words, an inferred synset was considered wrong if it had been falsely merged or if it had falsely been kept separate from another one. Figure 1 shows the precision and recall curves. All refinement methods improve precision, whereas recall only decreases at low thresholds. Figure 2 shows $F_1$ and $F_{0.5}$ scores; we deem the latter more sensible in the present setting as high precision is more important to us than high recall. The FT-C and SW2V methods perform best at lower thresholds, while Nasari works best at high thresholds.

An additional manual evaluation was carried out with 50 random German words[3] and four settings that obtained high $F_1$ or $F_{0.5}$ scores. As shown in Table 3, the SW2V method with a threshold set at 0.3 obtained the highest precision value by a large margin and therefore also the best $F_{0.5}$ score. We chose this setting for all languages. Source words that end up with a single synset as a result of this step are discarded.

[2]For both embeddings, we use the pre-trained 300-dimensional Babel synset representation trained on the UMBC corpus.

[3]All words were associated with at least two synsets by the baseline model, but only 18 out of them (36%) contained two or more synsets according to a human annotator.

| Method | Threshold | Prec. | Rec. | $F_1$ | $F_{0.5}$ |
|---|---|---|---|---|---|
| Baseline | | 33% | **48%** | 39% | 35% |
| Nasari | 0.55 | 54% | 31% | 39% | 47% |
| SW2V | 0.3 | **67%** | 28% | 40% | **52%** |
| | 0.5 | 46% | 42% | **44%** | 45% |
| FT-C | 0.35 | 54% | 27% | 36% | 45% |
| | 0.45 | 50% | 35% | 41% | 46% |

Table 3: Manual evaluation results for selected parameter settings.

### 2.4 Step 3: Selecting sentences and creating contrast pairs (Scoring variant only)

We use the synset lexicon built in the previous step to guide the creation of contrast pairs. We extract sentence pairs from the parallel corpora and group them by source word and target word sense. We restrict the extraction process to sentences longer than 10 words and skip sentences in which the source or target item occurs more than once. From this set, we randomly choose 20 instances of each sense from various corpus sources.

For each extracted sentence pair, a contrastive sentence pair is produced by keeping the source sentence identical, but replacing the target word in the target sentence by another lexicalisation from a different synset.

While this entirely automatic setup could give rise to inconsistencies which would require manual correction as in Rios Gonzales et al. (2017), we argue that BabelNet constraints already provide some filtering (for example mostly keeping number constant). Given our aim to scale up to a large number of languages, the need for human intervention would make the creation of a large scale multilingual benchmark difficult and costly.

### 2.5 Statistics

We apply the three steps presented above to all to-English translation directions that were part of the Conference of Machine Translation (WMT) news translation task over the last years. Table 4 summarizes the statistics of these resources. The average number of senses per source word ranges between 2.0 and 2.11 (2.36–2.4 for ContraWSD). The lexicons for the Baltic languages are small due to the small size of available parallel corpora.

## 3 Measuring machine translation WSD capability with MuCoW

The aim of MuCoW is to examine the ability of current machine translation systems to choose the correct target sense of ambiguous source words. Here, we give some baseline results obtained with supervised NMT systems. Following Rios Gonzales et al. (2017), we score both reference and contrastive translations with the same NMT system. A correct decision is detected when the score of the reference is higher than the scores from all contrastive translations. The final test suite score corresponds to the accuracy over all decisions.

Three models are examined for German→English: a 6-layer bi-LSTM model and a Transformer model[4] trained on the provided training data from WMT17 plus backtranslations from Sennrich et al. (2016b), and the University of Edinburgh's WMT17 submission, a deep LSTM model with additional synthetic data trained with Nematus (Sennrich et al., 2017b).[5] The upper half of Table 5 reports ContraWSD

| | Corpus | Lexicon | | | Test suite |
|---|---|---|---|---|---|
| Language pair | Sentence pairs | Source words | Target synsets | Target words | Sentence pairs |
| CS–EN | 44M | 107 | 223 | 412 | 11470 |
| DE–EN | 35M | 259 | 548 | 1086 | 33077 |
| ES–EN | 81M | 515 | 1090 | 2398 | 72295 |
| ET–EN | 14M | 34 | 68 | 89 | 2500 |
| FI–EN | 31M | 176 | 367 | 610 | 16326 |
| FR–EN | 68M | 456 | 963 | 2152 | 64369 |
| LT–EN | 2.5M | 10 | 20 | 31 | 922 |
| LV–EN | 1.6M | 5 | 10 | 12 | 318 |
| RO–EN | 52M | 129 | 263 | 496 | 14258 |
| RU–EN | 38M | 113 | 234 | 396 | 12378 |
| TR–EN | 46M | 107 | 220 | 420 | 11795 |

Table 4: Sizes of the parallel corpora used for lexicon extraction, the inferred and filtered ambiguity lexicons, and the resulting test suite corpora.

| Lg. pair | Model | ContraWSD | MuCoW | BLEU |
|---|---|---|---|---|
| DE–EN | LSTM | 77.55 | 60.50 | 30.3 |
| | Transformer | 86.42 | 66.98 | 33.3 |
| | Nematus | 86.72 | 68.80 | 35.1 |
| CS–EN | Nematus | | 78.77 | 30.9 |
| RO–EN | Nematus | | 62.86 | 33.3 |
| RU–EN | Nematus | | 72.36 | 30.8 |
| TR–EN | Nematus | | 62.69 | 20.1 |

Table 5: Comparison of MuCoW and ContraWSD accuracy scores and BLEU scores computed on the WMT news2017 test set (news2016 for RO–EN).

[4] Sentences are encoded using Byte-Pair Encoding (Sennrich et al., 2016c), with 32,000 merge operations for each language. For the Bi-LSTM model we use embedding layers and hidden units of 512 dimensions. For the Transformer, we use the *base* version (Vaswani et al., 2017).

[5] data.statmt.org/wmt17_systems/

and MuCoW accuracy scores as well as BLEU scores computed on the WMT17 test set. The ranking of the three models is consistent across the three tasks. Interestingly, the Transformer model (trained on far less data than the Nematus model) scores much better on the two test suites than the BLEU score would suggest, confirming the findings by Tang et al. (2018).

The University of Edinburgh also makes available their NMT models for other WMT16 and WMT17 language pairs.[6] MuCoW accuracy scores of these models are shown in the lower half of Table 5 together with the WMT test set BLEU scores reported by the authors (Sennrich et al., 2016a, 2017a).

Even though we only assess the confidence of an NMT system in detecting the right sense of a single word within a sentence, the results show that WSD is still an issue in MT – even in state-of-the-art-systems – that requires further study.

## 4 Translation test suites for WMT 2019

As mentioned in Section 1, the WMT test suite call requires a different setup that does not rely on scoring capabilities of the participating systems. Therefore, we modified step (3) of our method to conform with these requirements, analogously to the modification of ContraWSD by Rios et al. (2018). As a beneficial side effect, we were also able to include language pairs with non-English target languages.[7] The changes to step (3) are the following:

- The sentence pairs were filtered more aggressively. We only kept sentence pairs in which both the source and target words were tagged as NOUNs by the respective UDPipe part-of-speech tagger (Straka and Straková, 2017).
- Source sentences stemming from one of the WMT training corpora were excluded. We only used sentences from the following OPUS corpora: *Books*, *Tatoeba*, *TED2013*, *EUBookstore* and *OpenSubtitles2018*.
- We only kept synsets for which we found at least 4 example sentences, and we retained at most 10 example sentences per sense.
- If as a result of the above filters, all but one senses of a source word were removed, we removed the source word entirely.
- We distinguished between in-domain and out-of-domain synsets. A synset is considered out-of-domain if more than half of its example sentences come from *OpenSubtitles2018*. The intuition behind this distinction is that most participating systems will be tuned towards the news domain and thus will not handle features of colloquial speech reliably.
- We disregarded the automatically generated contrastive sentences.

| Language pair | Source words | Target synsets | In-dom synsets | Out-dom synsets | Sentences |
|---|---|---|---|---|---|
| DE–EN | 217 | 461 | 329 | 132 | 4268 |
| FI–EN | 109 | 231 | 91 | 140 | 2117 |
| LT–EN | 6 | 12 | 5 | 7 | 99 |
| RU–EN | 67 | 138 | 59 | 79 | 1223 |
| EN–CS | 98 | 200 | 29 | 171 | 1843 |
| EN–DE | 176 | 362 | 220 | 142 | 3337 |
| EN–FI | 48 | 97 | 22 | 75 | 830 |
| EN–LT | 4 | 8 | 3 | 5 | 69 |
| EN–RU | 97 | 199 | 40 | 163 | 1814 |

Table 6: Sizes of the MuCoW data sets compiled for WMT19.

We built the translation variant of MuCoW for 9 translation directions of the news task. Table 6 shows some statistics.

The resulting test suites contain sentences of the source language together with the following metadata: the ambiguous source word, the list of correct target words (the correct target synset), the list of incorrect target words (the incorrect target synset), and information about the domain of the synsets. Table 7 shows an example. The source language sentences were sent (without metadata) to the WMT participants as part of the test set, and we received the translations for evaluation.

## 5 WMT 2019 test suite results

In order to assess the translation output of the WMT participants, we check whether any of the correct or incorrect target words listed in the metadata file can be identified in the tokenized and lowercased translation output.

Although the sentences have been selected to contain the uninflected base form both in the

[6] data.statmt.org/wmt{16,17}_systems/

[7] We limited our work to from-English language pairs due to time restrictions, but the method would be generic enough to also work for French–German, German–French, and German–Czech.

| Example containing **ambiguous word** | Correct translations | Incorrect translations |
|---|---|---|
| It occurred to me that my **watch** might be broken. | Armbanduhr, Uhr | *Wache* |
| I hope you didn't get distracted during your **watch**. | *Wache* | Armbanduhr, Uhr |
| In winter, the dry leaves fly around in the **air**. | Luft, Luftraum, Aura | Miene, Ausdruck |
| He remained silent for a moment, with a thoughtful but contented **air**. | Miene, Ausdruck | Luft, Luftraum, Aura |
| Harry had to back out of the competition because of a broken **arm**. | Arm | *Waffe* |
| So does the cop who left his side arm in a subway bathroom. | *Waffe* | Arm |
| Drain the pasta and return the pasta to the **pot**. | Blumentopf, Kochtopf, Topf, Nachttopf | *Marihuana, Gras* |
| Where did those idiots get all of this **pot** anyhow? | *Marihuana, Gras* | Blumentopf, Kochtopf, Topf, Nachttopf |

Table 7: Examples of test suite instances of the English–German WMT test suite. The ambiguous (English) source word is highlighted in bold, and correct and incorrect (German) translations – as inferred by the MuCoW procedure – are given. Senses classified as out-of-domain are shown in italics. Note that some example sentences may further restrict the set of correct translations.

| Language pair | Average coverage (tokenized) | Average coverage (with lemma backoff) |
|---|---|---|
| DE–EN | 83.06% | 84.51% |
| FI–EN | 81.52% | 82.14% |
| LT–EN | 92.75% | 93.48% |
| RU–EN | 82.23% | 82.85% |
| EN–CS | 61.77% | 74.87% |
| EN–DE | 66.52% | 69.26% |
| EN–FI | 52.27% | 67.55% |
| EN–LT | 64.86% | 79.71% |
| EN–RU | 58.88% | 73.29% |

Table 8: Average coverage of target words among WMT19 primary submissions.

source and target languages, we cannot assume that all translation systems will output base forms. Hence, if neither correct nor incorrect target words can be identified, we lemmatize the translation output and search the target words again in the lemmatized version.[8] Depending on the target language, lemmatization allowed us to substantially increase the coverage (see Table 8).

We report precision, recall and F1-score for in-domain senses and out-of-domain senses, except for Lithuanian, where not enough examples are available. Precision and recall are computed as follows:[9]

$$\text{Precision} = \frac{\text{\# examples with correct target words}}{\text{\# examples with either correct or incorrect target words}}$$

$$\text{Recall} = \frac{\text{\# examples with correct target words}}{\text{\# total examples}}$$

For each language pair, EN→CS, EN↔DE, EN↔FI, EN↔RU and EN↔LT, results are shown respectively in Tables 9 to 13. Overall, we observe that systems perform quite well in WSD, achieving high precision overall. For some translation directions, there is a big gap between in-domain and out-of-domain synsets, showing clearly that systems tuned towards news translation struggle to identify the right sense when tested on a different domain. At the same time, online systems are more robust to domain mismatch, which is likely due to their use of a much larger variety of training data. Interestingly, the Czech–English task shows opposite results, with online systems performing better on in-domain synsets than research systems.

Interestingly enough, having English as source side yields better overall precision comparing with English as target side. One possible explanation could be found in the difficulty to obtain better encoder representations for morphologically rich languages. Recall is better with English on the target side due to higher coverage (Table 8).

It would have been instructive to compare the MuCoW results with automatic or manual evaluation scores on the official WMT19 test set, but unfortunately, such scores were not available in time for all systems.

## 6 Conclusion

In this paper, we have presented MuCoW, an automatically built WSD test suite for machine translation that relies on large parallel corpora, the multilingual lexical resource BabelNet and language-

[8] We used the Turku neural lemmatizer with pretrained models (Kanerva et al., 2019). For Lithuanian, as no pretrained model was available, we trained one using the respective available data from the Universal Dependencies project.

[9] Examples that contained both correct and incorrect target words were counted as incorrect.

independent synset embeddings. We used the proposed benchmark to assess the WSD ability of NMT systems following two evaluation protocols: scoring both reference and contrastive translations with pretrained NMT models, and as translation test suite for the WMT19 news shared task.

We find that state-of-the-art and fine-tuned NMT systems still present some drawbacks on handling ambiguous words, especially when evaluated on out-of-domain data and when the encoder has to deal with a morphologically rich language. It will be particularly instructive to see how well the WSD test suite results correlate with human evaluation scores and with recently proposed evaluation metrics that are based on semantic representations of the translations (Gupta et al., 2015; Shimanaka et al., 2018).

As future work we plan to further extend the test suite including more languages and parallel data, and make use of the contrastive sentences as adversarial examples during training.

## Acknowledgments


This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

The authors gratefully acknowledge the support of the Academy of Finland through project 314062 from the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence. Finally, We would also like to acknowledge NVIDIA and their GPU grant.


## References


Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A linguistic evaluation of rule-based, phrase-based, and neural MT engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th iwslt evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation, Heidelberg, Germany*.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA).

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 636–643. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal. Association for Computational Linguistics.

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, , Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes,

| Submission | In-domain synsets | | | Out-of-domain synsets | | | All synsets | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Prec. | Recall | F1 | Prec. | Recall | F1 |
| CUNI-Trf-T2T-2018 | 96.76 | 84.75 | 90.36 | **79.85** | **71.71** | **75.56** | **82.77** | 74.01 | **78.15** |
| CUNI-Trf-T2T-2019 | 95.60 | 85.66 | 90.36 | 79.58 | 71.57 | 75.36 | 82.38 | **74.04** | 77.99 |
| CUNI-DocTrf-T2T | 95.60 | 85.66 | 90.36 | 79.58 | 71.57 | 75.36 | 82.38 | **74.04** | 77.99 |
| CUNI-DocTrf-Marian | 96.00 | 85.71 | 90.57 | 72.45 | 68.51 | 70.42 | 76.61 | 71.69 | 74.07 |
| uedin | 96.30 | 83.27 | 89.31 | 72.96 | 67.85 | 70.31 | 77.02 | 70.70 | 73.72 |
| online-Y | 97.57 | 84.86 | 90.77 | 61.57 | 63.73 | 62.63 | 67.93 | 68.03 | 67.98 |
| parfda | 95.02 | 75.27 | 84.00 | 68.16 | 58.44 | 62.93 | 72.85 | 61.57 | 66.74 |
| online-X | 95.70 | **87.81** | **91.59** | 57.35 | 58.89 | 58.11 | 64.54 | 64.83 | 64.68 |
| online-A | 95.88 | 83.21 | 89.10 | 58.36 | 58.25 | 58.30 | 65.17 | 63.33 | 64.24 |
| online-B | **97.93** | 83.16 | 89.94 | 57.02 | 57.24 | 57.13 | 64.46 | 62.63 | 63.53 |

Table 9: Results for English–Czech.

| Submission | In-domain synsets | | | Out-of-domain synsets | | | All synsets | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Prec. | Recall | F1 | Prec. | Recall | F1 |
| **German–English:** | | | | | | | | | |
| Facebook_FAIR | **80.78** | **85.80** | **83.21** | **52.77** | **72.56** | **61.10** | **73.55** | **82.99** | **77.99** |
| online-B | 77.88 | 83.81 | 80.73 | 45.50 | 66.51 | 54.04 | 69.58 | 80.31 | 74.56 |
| online-G | 77.62 | 83.76 | 80.57 | 45.62 | 65.43 | 53.76 | 69.48 | 80.02 | 74.38 |
| online-Y | 76.82 | 84.51 | 80.48 | 41.93 | 61.71 | 49.93 | 68.10 | 79.97 | 73.56 |
| dfki-nmt | 77.64 | 83.35 | 80.39 | 41.08 | 63.02 | 49.74 | 68.31 | 79.42 | 73.45 |
| RWTH_Aachen | 77.62 | 84.30 | 80.83 | 36.96 | 60.92 | 46.01 | 67.30 | 80.02 | 73.11 |
| MSRA.MADL | 77.95 | 84.36 | 81.03 | 36.73 | 56.26 | 44.44 | 67.78 | 79.08 | 73.00 |
| UCAM | 76.79 | 84.04 | 80.25 | 35.38 | 55.71 | 43.28 | 66.54 | 78.77 | 72.14 |
| MLLP-UPV | 77.26 | 83.24 | 80.14 | 35.85 | 54.92 | 43.38 | 67.02 | 77.93 | 72.06 |
| online-A | 75.77 | 83.08 | 79.26 | 37.47 | 63.15 | 47.04 | 65.87 | 79.40 | 72.00 |
| NEU | 75.26 | 83.50 | 79.16 | 32.49 | 55.93 | 41.11 | 64.49 | 78.58 | 70.84 |
| JHU | 74.94 | 83.68 | 79.07 | 31.56 | 51.38 | 39.10 | 64.31 | 77.79 | 70.41 |
| uedin | 74.26 | 81.62 | 77.77 | 32.21 | 45.89 | 37.85 | 64.28 | 74.70 | 69.10 |
| PROMT_NMT | 70.05 | 81.34 | 75.27 | 32.02 | 43.94 | 37.05 | 61.20 | 73.70 | 66.87 |
| online-X | 67.04 | 80.29 | 73.07 | 31.98 | 62.47 | 42.31 | 57.77 | 77.07 | 66.04 |
| TartuNLP-c | 71.11 | 77.22 | 74.04 | 29.29 | 46.31 | 35.88 | 60.68 | 71.48 | 65.64 |
| **English–German:** | | | | | | | | | |
| Facebook_FAIR | **83.43** | 76.99 | **80.08** | **56.29** | **55.10** | **55.69** | **74.48** | **70.05** | **72.19** |
| Microsoft-sentence-level | 83.18 | **77.14** | 80.05 | 52.81 | 51.92 | 52.36 | 73.31 | 69.27 | 71.23 |
| online-B | 83.37 | 74.78 | 78.85 | 51.92 | 50.66 | 51.28 | 73.04 | 67.30 | 70.05 |
| Microsoft-document-level | 81.76 | 75.68 | 78.60 | 47.21 | 48.11 | 47.65 | 70.54 | 67.29 | 68.88 |
| online-Y | 81.29 | 75.30 | 78.18 | 46.37 | 48.21 | 47.27 | 69.87 | 67.12 | 68.47 |
| online-G | 81.44 | 73.76 | 77.41 | 46.61 | 45.44 | 46.02 | 70.21 | 65.09 | 67.55 |
| dfki-nmt | 80.70 | 74.37 | 77.41 | 44.95 | 42.04 | 43.44 | 69.54 | 64.39 | 66.87 |
| MLLP-UPV | 79.90 | 73.60 | 76.62 | 44.03 | 39.63 | 41.72 | 68.90 | 63.01 | 65.82 |
| lmu-ctx-tf-single | 79.55 | 72.51 | 75.86 | 43.93 | 41.99 | 42.94 | 68.23 | 63.13 | 65.58 |
| NEU | 78.39 | 73.50 | 75.86 | 41.91 | 41.53 | 41.72 | 66.83 | 63.75 | 65.25 |
| eTranslation | 80.44 | 71.00 | 75.43 | 43.47 | 40.48 | 41.92 | 68.69 | 61.65 | 64.98 |
| MSRA.MADL | 80.53 | 71.97 | 76.01 | 41.79 | 35.63 | 38.46 | 68.88 | 60.67 | 64.51 |
| UCAM | 78.21 | 72.70 | 75.35 | 40.41 | 37.28 | 38.78 | 66.61 | 61.77 | 64.10 |
| online-A | 79.21 | 72.05 | 75.46 | 40.48 | 36.44 | 38.35 | 67.37 | 61.09 | 64.07 |
| Helsinki-NLP | 78.34 | 72.52 | 75.32 | 39.06 | 36.65 | 37.82 | 66.24 | 61.57 | 63.82 |
| PROMT_NMT | 78.08 | 72.40 | 75.13 | 36.99 | 34.16 | 35.52 | 65.61 | 60.77 | 63.10 |
| JHU | 77.80 | 71.48 | 74.50 | 37.77 | 29.35 | 33.04 | 66.47 | 58.08 | 61.99 |
| UdS-DFKI | 78.27 | 70.54 | 74.21 | 35.68 | 30.16 | 32.69 | 65.72 | 58.10 | 61.68 |
| online-X | 71.01 | 72.71 | 71.85 | 34.36 | 40.47 | 37.17 | 59.07 | 63.16 | 61.05 |
| TartuNLP-c | 77.32 | 66.29 | 71.38 | 33.02 | 26.13 | 29.17 | 64.34 | 53.85 | 58.63 |
| en_de_task | 64.54 | 23.14 | 34.06 | 38.41 | 5.64 | 9.84 | 59.43 | 16.62 | 25.97 |

Table 10: Results for German–English and English–German.

| Submission | In-domain synsets | | | Out-of-domain synsets | | | All synsets | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Prec. | Recall | F1 | Prec. | Recall | F1 |
| **Finnish–English:** | | | | | | | | | |
| online-G | 78.00 | 84.17 | 80.97 | **71.47** | 81.65 | **76.22** | **74.14** | **82.71** | **78.19** |
| online-Y | 79.30 | 82.89 | 81.05 | 63.40 | **81.73** | 71.41 | 69.78 | 82.25 | 75.51 |
| GTCOM-Primary | 81.87 | **84.81** | **83.31** | 57.28 | 77.64 | 65.92 | 67.36 | 81.05 | 73.57 |
| MSRA.NAO | **82.21** | 83.79 | 82.99 | 57.26 | 77.86 | 65.99 | 67.42 | 80.70 | 73.46 |
| USYD | 80.05 | 83.43 | 81.71 | 56.18 | 71.50 | 62.92 | 66.20 | 77.09 | 71.23 |
| parfda | 77.89 | 78.66 | 78.27 | 55.16 | 66.01 | 60.10 | 64.71 | 71.86 | 68.10 |
| online-B | 77.55 | 82.01 | 79.72 | 52.10 | 66.97 | 58.61 | 62.88 | 74.07 | 68.02 |
| online-A | 76.16 | 78.70 | 77.41 | 52.85 | 69.02 | 59.87 | 62.46 | 73.57 | 67.56 |
| Helsinki-NLP | 76.65 | 78.53 | 77.58 | 48.52 | 62.86 | 54.77 | 60.37 | 70.37 | 64.99 |
| online-X | 68.92 | 76.68 | 72.59 | 51.39 | 67.75 | 58.45 | 58.63 | 71.81 | 64.56 |
| TartuNLP-c | 75.35 | 79.77 | 77.49 | 45.32 | 53.13 | 48.92 | 58.70 | 65.68 | 61.99 |
| apertium-unconstrained | 63.97 | 67.15 | 65.52 | 38.46 | 52.86 | 44.53 | 48.96 | 59.69 | 53.80 |
| **English–Finnish:** | | | | | | | | | |
| online-G | 93.71 | 75.25 | 83.47 | **80.62** | **68.54** | **74.09** | **84.01** | **70.36** | **76.58** |
| online-Y | 94.74 | 72.00 | 81.82 | 75.06 | 66.08 | 70.28 | 80.03 | 67.75 | 73.38 |
| MSRA.NAO | **95.62** | **76.12** | **84.76** | 68.47 | 66.60 | 67.52 | 75.44 | 69.42 | 72.31 |
| GTCOM-Primary | 94.81 | 73.00 | 82.49 | 66.24 | 67.97 | 67.09 | 73.25 | 69.49 | 71.32 |
| online-X | 84.14 | 65.95 | 73.94 | 62.22 | 61.95 | 62.08 | 67.56 | 63.11 | 65.26 |
| NICT | 90.32 | 72.54 | 80.46 | 57.62 | 59.35 | 58.48 | 66.06 | 63.42 | 64.71 |
| online-B | 88.75 | 74.74 | 81.14 | 59.02 | 56.38 | 57.67 | 67.12 | 61.85 | 64.38 |
| Aalto-ORMFC | 88.81 | 66.15 | 75.82 | 64.94 | 54.79 | 59.44 | 71.17 | 58.04 | 63.93 |
| Helsinki-NLP | 84.56 | 61.50 | 71.21 | 59.65 | 52.51 | 55.85 | 65.93 | 55.11 | 60.03 |
| online-A | 86.75 | 77.42 | 81.82 | 52.31 | 46.79 | 49.39 | 62.59 | 55.95 | 59.08 |
| TartuNLP-c | 93.29 | 70.20 | 80.12 | 53.83 | 43.49 | 48.11 | 65.24 | 51.61 | 57.63 |
| Helsinki-NLP-rule-based | 71.60 | 75.62 | 73.56 | 48.88 | 47.36 | 48.11 | 55.59 | 55.21 | 55.40 |
| apertium-unconstrained | 81.71 | 34.72 | 48.73 | 45.61 | 20.88 | 28.65 | 55.16 | 24.75 | 34.17 |

Table 11: Results for Finnish–English and English–Finnish.

| Submission | In-domain synsets | | | Out-of-domain synsets | | | All synsets | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Prec. | Recall | F1 | Prec. | Recall | F1 |
| **Russian–English:** | | | | | | | | | |
| online-G | **92.15** | 89.63 | **90.87** | **66.95** | **80.87** | **73.26** | **78.57** | **85.38** | **81.84** |
| Facebook_FAIR | 89.98 | **89.80** | 89.89 | 56.67 | 77.30 | 65.40 | 72.12 | 84.07 | 77.64 |
| online-B | 89.55 | 87.58 | 88.55 | 56.41 | 74.07 | 64.04 | 71.81 | 81.34 | 76.28 |
| online-A | 87.93 | 87.58 | 87.76 | 50.97 | 73.16 | 60.08 | 68.09 | 81.15 | 74.05 |
| online-Y | 88.68 | 87.07 | 87.87 | 50.90 | 70.75 | 59.21 | 68.52 | 79.78 | 73.72 |
| MSRA.SCA | 86.22 | 85.33 | 85.77 | 50.27 | 72.45 | 59.35 | 66.76 | 79.57 | 72.60 |
| NEU | 87.19 | 86.48 | 86.83 | 47.89 | 72.15 | 57.57 | 65.97 | 80.23 | 72.40 |
| afrl-syscomb19 | 86.85 | 85.42 | 86.13 | 44.40 | 65.41 | 52.90 | 64.26 | 76.78 | 69.96 |
| eTranslation | 87.71 | 84.15 | 85.89 | 43.82 | 62.73 | 51.60 | 64.41 | 74.91 | 69.27 |
| rerank-re | 87.71 | 84.15 | 85.89 | 43.23 | 61.99 | 50.94 | 64.14 | 74.62 | 68.99 |
| online-X | 82.39 | 87.90 | 85.06 | 35.99 | 65.06 | 46.35 | 57.66 | 78.71 | 66.56 |
| TartuNLP-u | 84.11 | 87.50 | 85.77 | 37.35 | 53.09 | 43.85 | 60.38 | 72.71 | 65.97 |
| afrl-ewc | 87.04 | 82.24 | 84.58 | 33.75 | 45.63 | 38.80 | 59.92 | 66.86 | 63.20 |
| NICT | 78.62 | 69.11 | 73.56 | 30.17 | 24.42 | 26.99 | 56.29 | 47.59 | 51.58 |
| **English–Russian:** | | | | | | | | | |
| online-G | **95.56** | 89.58 | 92.47 | **75.11** | **74.85** | **74.98** | **80.05** | **78.58** | **79.31** |
| Facebook_FAIR | 95.49 | 88.28 | 91.75 | 67.68 | 71.54 | 69.56 | 74.40 | 76.01 | 75.20 |
| online-B | 95.08 | 91.10 | 93.05 | 62.12 | 69.05 | 65.40 | 70.31 | 75.16 | 72.66 |
| USTC-MCC | 95.30 | 90.08 | 92.62 | 59.35 | 71.08 | 64.69 | 68.02 | 76.54 | 72.03 |
| NEU | 94.43 | 89.21 | 91.75 | 59.31 | 70.98 | 64.62 | 67.74 | 76.18 | 71.71 |
| online-Y | 95.37 | 91.38 | **93.33** | 57.47 | 69.02 | 62.72 | 66.80 | 75.51 | 70.89 |
| online-A | 91.14 | 89.40 | 90.26 | 55.29 | 68.28 | 61.10 | 64.00 | 74.35 | 68.79 |
| PROMT_NMT | 93.48 | **91.49** | 92.47 | 56.78 | 63.76 | 60.07 | 66.18 | 71.61 | 68.79 |
| online-X | 93.65 | 89.92 | 91.75 | 52.53 | 67.35 | 59.02 | 62.53 | 74.12 | 67.83 |
| TartuNLP-u | 90.91 | 84.01 | 87.32 | 51.44 | 56.17 | 53.70 | 61.41 | 64.11 | 62.73 |
| rerank-er | 94.98 | 78.91 | 86.20 | 55.54 | 33.78 | 42.01 | 68.17 | 45.36 | 54.47 |
| NICT | 89.19 | 25.52 | 39.68 | 46.99 | 5.88 | 10.46 | 63.90 | 10.33 | 17.78 |

Table 12: Results for Russian–English and English–Russian.

| Submission | All synsets Prec. | Recall | F1 | Submission | All synsets Prec. | Recall | F1 |
|---|---|---|---|---|---|---|---|
| **Lithuanian–English:** | | | | **English–Lithuanian:** | | | |
| tilde-c-nmt | **80.41** | 97.50 | **88.14** | MSRA.MASS | 78.69 | **85.71** | **82.05** |
| NEU | 79.59 | **98.73** | **88.14** | online-B | 79.31 | 80.70 | 80.00 |
| tilde-nc-nmt | 79.38 | 97.47 | 87.50 | tilde-nc-nmt | 80.70 | 79.31 | 80.00 |
| GTCOM-Primary | 77.32 | 97.40 | 86.21 | tilde-c-nmt | 81.82 | 76.27 | 78.95 |
| online-B | 75.51 | 98.67 | 85.55 | MSRA.MASS | 78.95 | 78.95 | 78.95 |
| MSRA.MASS | 73.47 | 98.63 | 84.21 | online-A | **83.02** | 73.33 | 77.88 |
| online-A | 73.96 | 95.95 | 83.53 | GTCOM-Primary | 78.57 | 77.19 | 77.88 |
| online-G | 72.92 | 95.89 | 82.84 | NEU | 76.79 | 76.79 | 76.79 |
| online-X | 60.22 | 90.32 | 72.26 | eTranslation | 79.25 | 72.41 | 75.68 |
| JUMT | 71.62 | 67.95 | 69.74 | TartuNLP-c | 81.25 | 65.00 | 72.22 |
| TartuNLP-c | 64.86 | 65.75 | 65.31 | online-X | 70.37 | 71.70 | 71.03 |
| | | | | online-G | 71.15 | 68.52 | 69.81 |

Table 13: Results for Lithuanian–English and English–Lithuanian.


Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. ArXiv:1803.05567.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2019. Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. *arXiv preprint arXiv:1902.00972*.

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

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 1336–1345.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2017. Embedding words and senses together via joint knowledge-enhanced training. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 100–111, Vancouver, Canada. Association for Computational Linguistics.

Rebecca Marvin and Philipp Koehn. 2018. Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, volume 1, pages 125–131.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72. Association for Computational Linguistics.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

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: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. The University of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399. Association for Computational Linguistics.

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. 2017b. 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. Association for Computational Linguistics.

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: Volume 2, Shared Task Papers*, pages 371–376. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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)*, volume 1, pages 1715–1725.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123. Association for Computational Linguistics.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.