An Evaluation Benchmark for Testing the Word Sense Disambiguation Capabilities of Machine Translation Systems

Alessandro Raganato, Yves Scherrer, Jörg Tiedemann
University of Helsinki
Department of Digital Humanities, Faculty of Arts
{alessandro.raganato,yves.scherrer,jorg.tiedemann}@helsinki.fi

Abstract

Lexical ambiguity is one of the many challenging linguistic phenomena involved in translation, i.e., translating an ambiguous word with its correct sense. In this respect, previous work has shown that the translation quality of neural machine translation systems can be improved by explicitly modelling the senses of ambiguous words. Recently, several evaluation test sets have been proposed to measure the word sense disambiguation (WSD) capability of machine translation systems. However, to date, these evaluation test sets do not include any training data that would provide a fair setup measuring the sense distributions present within the training data itself. In this paper, we present an evaluation benchmark on WSD for machine translation for 10 language pairs, comprising training data with known sense distributions. Our approach for the construction of the benchmark builds upon the wide-coverage multilingual sense inventory of BabelNet, the multilingual neural parsing pipeline TurkuNLP, and the OPUS collection of translated texts from the web. The test suite is available at `http://github.com/Helsinki-NLP/MuCoW`.

Keywords: machine translation, word sense disambiguation, evaluation benchmark, test suite

1. Introduction

In recent years, several advances have been made in Word Sense Disambiguation (WSD) (Raganato et al., 2017; Loureiro and Jorge, 2019; Kumar et al., 2019). WSD models that tackle lexical ambiguity effectively bring numerous benefits to a variety of downstream tasks and applications, such as information retrieval and extraction (Zhong and Ng, 2012; Delli Bovi et al., 2015) and text categorization (Flekova and Gurevych, 2016; Pilehvar et al., 2017; Sinoara et al., 2019; Shimura et al., 2019). Another downstream application is machine translation (MT), where word sense disambiguation plays a crucial role to select the correct translation sense for each ambiguous word (Rios et al., 2017; Pu et al., 2018; Liu et al., 2018).

To measure specific linguistic phenomena in machine translation, several test suites – or challenge sets – have emerged (Popović and Castilho, 2019). They are evaluation benchmarks that focus on particular linguistic phenomena and provide specific evaluation criteria or metrics. For lexical ambiguity of nouns, to our knowledge, two major test suites exist: ContraWSD (Rios et al., 2017; Rios et al., 2018) and MuCoW (Raganato et al., 2019). Both test suites are available in two variants: *scoring* and *translation*. The first variant relies on the ability of neural machine translation systems to score given translations: 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. A contrast is considered successfully detected if the reference translation obtains a higher score than the artificially modified translation. The second variant relies directly on the translation produced by the system. After a system translates a sentence containing an ambiguous source word, an evaluation script checks whether any of the correct or incorrect target senses can be identified in the translation output. While both variants have different pros and cons, the translation one allows an evaluation directly on the output of a system, avoiding the need for a function for scoring a translation, which is typically not available for online systems or unsupervised MT approaches.

Recent works suggest that the state-of-the-art Transformer architecture (Vaswani et al., 2017) for neural MT (NMT) is able to deal with lexical ambiguity quite well (Tang et al., 2018; Tang et al., 2019), learning to distinguish between senses during translation with high precision. Prior works have shown that NMT models based on Recurrent Neural Networks (RNNs) struggle when dealing with rare word senses (Rios et al., 2017), but it is not fully clear how well the more recent Transformer architecture performs under different sense frequencies, size of training corpora and across different language pairs.

In earlier work (Raganato et al., 2019), we have presented MuCoW, a language-independent and fully automatic method for building a test suite for lexically ambiguous nouns. Here, we report on an extension of this work that covers the following aspects:

- We use an improved selection of data sources for building MuCoW to reduce noise and domain effects.
- The sense inference process is streamlined and relies on lemmatization instead of word alignment, leading to better coverage especially for morphologically rich languages.
- We provide not only the test sets, but also training sets. This guarantees that competing translation models are evaluated on fair grounds.
- We additionally provide an evaluation on the resulting benchmark with the state-of-the-art Transformer architecture.

We make available training and test sets for 10 language pairs (English ↔ Czech, English ↔ German, English

	CS–EN	DE–EN	FI–EN	FR–EN	RU–EN
Books		✓	✓	✓	✓
GlobalVoices	✓	✓		✓	✓
Europarl	✓	✓	✓	✓	
JW300	✓	✓	✓	✓	✓
News-Comm.	✓	✓		✓	✓
Tatoeba	✓	✓	✓	✓	✓
TED Talks		✓		✓	✓
EU Bookshop		✓	✓	✓	✓
MultiUN		✓		✓	✓
Common Crawl	✓	✓			✓

Table 1: Corpora used to extract the MuCoW test suites. The upper part defines the "clean" corpus sources, the lower part the "noisy" sources. The same corpora are used for both translation directions.

↔ Finnish, English ↔ French, and English ↔ Russian) with a total of 206 395 test sentences. The data and scoring scripts are available at `https://github.com/Helsinki-NLP/MuCoW`.

2. Methodology: Building MuCoW

The gist of our approach lies in the combination of different resources and tools: the wide-coverage multilingual sense inventory of BabelNet (Navigli and Ponzetto, 2012) and its associated sense embeddings (Mancini et al., 2017), the OPUS collection of translated texts from the web (Tiedemann, 2012), and the multilingual neural parsing pipeline TurkuNLP (Kanerva et al., 2018). In the following, we describe the three steps needed to create a MuCoW test suite.

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

For each language pair, we determine a set of parallel text sources mainly from the OPUS collection (see Table 1).[1] Based on our previous experience, we do not include movie subtitle data here to reduce the domain and genre variance. All corpora are tagged and lemmatized using the TurkuNLP neural parser pipeline with pre-trained models.[2]

Using the sentence-aligned corpora and the lexical resource BabelNet, we create lists of ambiguous nouns in the following way: if the source sentence contains a noun that is in BabelNet, and the target sentence contains a noun that is one of the translations of the source noun provided by BabelNet, we add the source noun together with the BabelNet sense id of the target noun to the list. We also keep track of the sentence pair. Since BabelNet mainly contains base forms, lemmatization is important to match as many sentences as possible.

[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, Tatoeba v2, TED2013 v1.1 (Cettolo et al., 2013), GlobalVoices v2017q3, JW300 (Agić and Vulić, 2019) and Common Crawl (web-crawled parallel corpus).

[2] All models are available online `http://bionlp-www.utu.fi/dep-parser-models/`. We used models_cs_pdt for Czech, models_en_ewt for English, models_fi_tdt for Finnish, models_fr_gsd for French, models_de_gsd for German, and models_ru_syntagrus for Russian.

Source word	BabelNet id	Target words
Krebs	bn:00023438n	brachyura, crab
	bn:00015180n	cancer (disease)
	bn:00015182n	Cancer (zodiac sign)
Quelle	bn:00080861n	well, borewell
	bn:00036077n	spring, fountain
	bn:00046702n	source, informant

Table 2: Examples of sense cluster refinement for German–English. In the first case (*Krebs*), the second and third clusters are merged because of the common target word *cancer*. In the second case (*Quelle*), the first and second clusters are merged because their similarity is above the threshold.

Sense ids for which less than 20 examples were found are discarded. Unambiguous source nouns, i.e., those associated with only one sense id, are discarded as well. We call each pairing of a source noun with a sense id *sense cluster*.

2.2. Step 2: 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 two additional merging steps:

1. We merge those sense clusters that share at least one common target word in BabelNet.
2. We merge sense clusters with similar meanings, as defined by their sense embeddings (Mancini et al., 2017). Concretely, following earlier work (Raganato et al., 2019), we merge senses whose cosine similarity is higher than 0.3.

Table 2 shows examples of both steps. Source nouns that become unambiguous as a result of this merging are again discarded.

2.3. Step 3: Selecting sentences for training and test sets

The sense lexicon built in the previous steps guides the selection of example sentences. We extract sentence pairs from the parallel corpora, using prioritarily data sources deemed as "clean" (see Table 1) and complementarily "noisy" data sources. Sentences in which the source or target item occurs more than once are skipped, and duplicates removed. The sentences are dispatched to the training and test sets as follows:

1. Add examples from the clean corpora to the test set until it contains 100 examples or 50% of all available examples.
2. Complete the test set with examples from the noisy corpora until it reaches 100 examples or 50% of all available examples.
3. Add the unused examples from the clean corpora to the training set.

Language pair	Ambiguous words	Sense clusters	Avg. sense similarity	Test sentences	Training sentences	Extended training sets **Small**	**Big**
CS–EN	27	54	0.185	4 465	150 710	0.6M	1.7M
DE–EN	135	281	0.245	24 688	414 499	1.4M	3.7M
FI–EN	60	122	0.174	10 260	263 890	1.4M	3.2M
FR–EN	219	457	0.250	38 609	995 002	1.8M	4.2M
RU–EN	50	103	0.203	8 657	88 351	1.1M	1.3M
EN–CS	87	178	0.242	13 229	211 260	0.4M	1.5M
EN–DE	206	429	0.351	35 315	688 499	1.2M	3.0M
EN–FI	128	261	0.369	21 481	582 755	1.0M	2.7M
EN–FR	231	473	0.380	39 860	993 723	1.6M	3.6M
EN–RU	58	120	0.374	9 831	98 483	0.6M	1.3M

Table 3: Sizes of the ambiguity lexicons and the training and test sets.

4. Complete the training set with examples from the noisy corpora until it contains at most 1000 examples.

Additionally, we make sure that the examples in the test set are at least 10 words long, and that the different corpora are equally represented.
The training sets obtained with this procedure are rather small, since they only contain sentences with ambiguous words. In order to provide realistic data conditions, we complete the training sets by adding sentence pairs that do not contain any ambiguous word in the source language from our lists. In the **Small** condition, we add all remaining sentences from Europarl only,[3] while in the **Big** condition, we add all remaining sentences from all clean corpora.
Thanks to this procedure, we know precisely how often each ambiguous word and sense is represented in the training data. Moreover, we can study the effect of increased target-independent training data on disambiguation performance.

2.4. Statistics

Table 3 shows the statistics of the sense lexicons and the training and test sets. It can be seen that the overwhelming majority of ambiguous words has two senses, in all language pairs. Furthermore, English as source language produces higher numbers of ambiguous words compared to English as target language. On the other hand, the extended training sets tend to be larger in the language pairs with English as source language.

3. Evaluation protocol

Together with the training and test data, we provide an evaluation script. Thanks to the controlled construction of training sets, different models and architectures can be compared fairly with each other. We suggest the following protocol:

1. Train the model exclusively with the provided training data.
2. Tune the model parameters to optimize general translation quality, e.g. by using the Newstest development sets provided by WMT.
3. Translate the provided test sentences.
4. Lemmatize the translated test sentences using the Turku neural parser pipeline.
5. Evaluate the translated test sentences with the provided evaluation script.[4]

The evaluation script identifies lemmas of the correct sense and lemmas of incorrect senses. On this basis, precision and recall are computed as shown below, with F1-score being computed in the usual way:[5]

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

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

These measures are broken down by source corpus, by relative frequency bucket of the sense, and by whether the source or target word has been segmented by a subword splitting method.
Furthermore, the script reports weighted precision based on the cosine similarity sim between the sense embedding of the correct sense and the embedding of the inferred sense. sim equals to 1 if the prediction is correct, and to 0 if it is maximally wrong:

$$\text{WeightedPrecision} = \frac{\sum sim(\text{correct sense, inferred sense})}{\text{\# examples with either correct or incorrect target lemmas}}$$

4. Baseline models

We have trained NMT models for all language pairs covered by MUCOW, using both extended training sets (**Small** and **Big**) as described above. We trained a 6-layer Transformer model (Vaswani et al., 2017)[6] for each language

[3] Except for Russian, where no Europarl data are available.

[4] More details about the evaluation protocol and the use of the evaluation script can be found on Github.

[5] Examples that contain both correct and incorrect target lemmas are counted as incorrect.

[6] As hyper-parameters we used the *base* version, with 100 000 training steps.

	Small					Big				
	Newstest	MuCoW				Newstest	MuCoW			
Lang. pair	BLEU	BLEU	Precision	Recall	F1-score	BLEU	BLEU	Precision	Recall	F1-score
CS–EN	21.48	27.97	0.7746	0.8852	0.8262	27.21	33.33	0.8044	0.9059	0.8522
DE–EN	24.43	24.53	0.7690	0.8649	0.8141	28.53	27.56	0.7831	0.8908	0.8334
FI–EN	21.60	26.02	0.8032	0.7991	0.8012	26.16	29.47	0.8125	0.8223	0.8174
FR–EN	29.68	26.52	0.7577	0.8380	0.7958	32.35	28.60	0.7668	0.8421	0.8027
RU–EN	24.30	21.93	0.7884	0.8661	0.8254	26.43	22.98	0.7894	0.8615	0.8239
EN–CS	15.20	21.89	0.7493	0.7716	0.7603	20.61	26.71	0.7859	0.8057	0.7957
EN–DE	19.75	20.39	0.7711	0.8015	0.7860	23.06	22.58	0.7829	0.8059	0.7942
EN–FI	16.40	18.10	0.7878	0.7440	0.7653	20.89	21.09	0.8003	0.7749	0.7874
EN–FR	27.53	24.01	0.7381	0.8240	0.7787	29.58	25.51	0.7303	0.8318	0.7778
EN–RU	18.88	16.99	0.7850	0.7635	0.7741	26.46	20.69	0.7965	0.8113	0.8038

Table 4: Results of the baseline models.

	Small					Big				
Lang. pair	0–20%	20–40%	40–60%	60–80%	80–100%	0–20%	20–40%	40–60%	60–80%	80–100%
CS–EN	0.7046	0.5233	0.7664	0.9464	0.9504	0.7078	**0.8075**	**0.8472**	0.9363	0.9525
DE–EN	0.5164	0.7293	0.8588	0.9129	0.9462	**0.5604**	**0.7704**	0.8680	0.9263	0.9547
FI–EN	0.6091	0.7162	0.7519	0.8949	0.9299	0.6156	0.7105	0.7733	0.9176	0.9498
FR–EN	0.4857	0.7407	0.8150	0.8647	0.9455	0.5091	0.7591	0.8118	0.8709	0.9474
RU–EN	0.5441	0.7280	0.8738	0.8575	0.9549	0.5512	0.7327	0.8667	0.8575	0.9505
EN–CS	0.4626	0.6752	0.7871	0.8054	0.9168	**0.5437**	**0.7320**	**0.8202**	**0.8370**	0.9296
EN–DE	0.4651	0.7047	0.8383	0.8352	0.9311	0.4701	0.7120	0.8525	0.8465	0.9364
EN–FI	0.5193	0.6977	0.7651	0.8315	0.8896	0.5444	0.7159	**0.7984**	0.8476	0.9099
EN–FR	0.4142	0.7070	0.8008	0.8704	0.9433	0.3888	0.6968	0.7998	0.8869	0.9480
EN–RU	0.4428	0.6982	0.8281	0.8811	0.9001	**0.4947**	**0.7546**	0.8516	0.8911	0.9239

Table 5: F1-scores of the baseline models, broken down by frequency bins. Bolded values indicate an improvement of at least 0.03 absolute compared to the other (Small or Big) model.

pair. Sentences are encoded using Truecaser and Byte-Pair Encoding (Sennrich et al., 2016), with 32 000 merge operations for each language, learned on each training corpus separately. Note that these models are not specifically adapted towards good sense disambiguation performance. They merely show how well off-the-shelf NMT architectures perform on lexical ambiguities. It is expected that architectures specifically adapted to WSD (Pu et al., 2018, for instance) would show substantially higher scores.

Table 4 summarizes the results of our experiments. The first column indicates general translation quality, as measured by BLEU[7] score on the Newstest corpora.[8] The **Big** training corpus shows substantial increases for all language pairs. Translating to English yields higher BLEU scores than translating to most other languages – a common finding in MT evaluations using BLEU.

Since MuCoW has been extracted from parallel corpora, we can also evaluate the test set with BLEU score. This is shown in the second column. Depending on the exact composition of the test set, its scores are lower or higher than those obtained on Newstest, but they are generally comparable.

BLEU measures overall translation quality and does not specifically focus on lexical ambiguity. This is measured by the three remaining columns, using precision, recall and F1-score (as defined in Section 3.). Overall, the absolute numbers are relatively high, with F1-scores ranging between 75% and 86%, suggesting that neural MT models can handle lexical ambiguity quite well out-of-the-box.

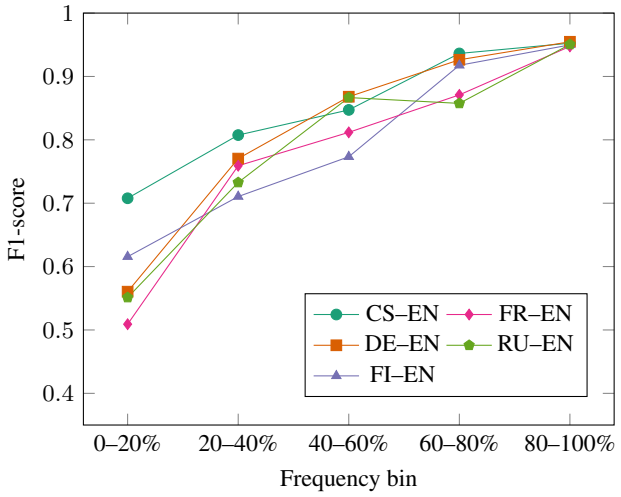
We see small, but consistent increases when moving from the **Small** to the **Big** setup. This is surprising – recall that the number of ambiguous words is identical in both training corpora – and suggests that the additional training data helps the model create more abstract and robust representations of words.

4.1. Frequency effects

Rare senses often are more difficult to translate, resulting in a dramatic drop in WSD capability (Rios et al., 2017). In order to quantify this, we classify the test sentences according to the frequency of the sense in the training corpus. For example, the "cancer" sense of German *Krebs* occurs in 98.1% of training examples, whereas the "crab" sense occurs in 1.9% of examples. We compute precision, recall

[7] We used sacreBLEU (Papineni et al., 2002; Post, 2018) with signature BLEU+case.lc+#.1+smooth.exp+tok.13a+v.1.2.11

[8] Newstest 2017 for English↔Finnish, and Newstest 2014 for the remaining language pairs.



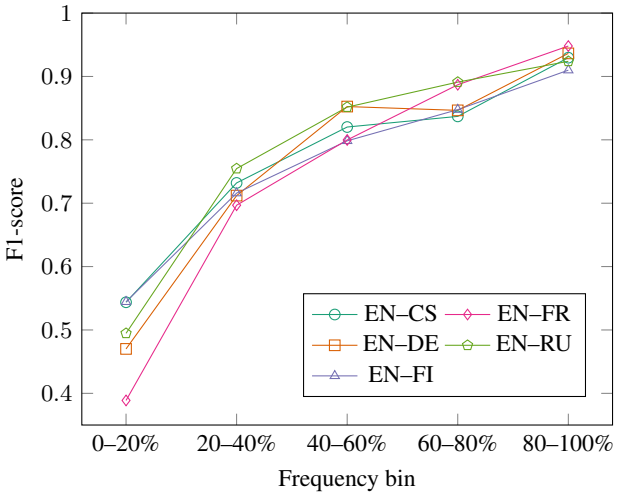


Figure 1: F1-scores for different relative frequency bins of sense clusters (to English on the left, from English on the right).

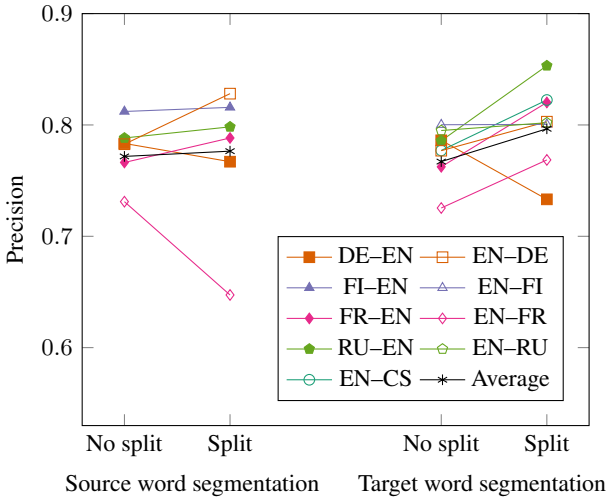


Figure 2: Precision values for different modes of word segmentation. Language pairs with unreliable figures (less than 50 examples per category) are removed.

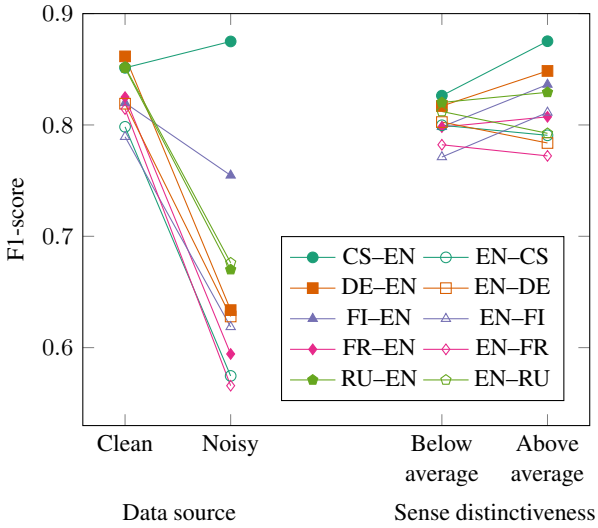


Figure 3: F1-scores broken down by provenience of test sentences (left) and by distinctiveness of the sense (right).

and F1-measure separately for five frequency bins. The results are listed in Table 5 and visualized in Figure 1 (**Big** models only).

As expected, the models show poor performance (F1-scores between 0.4 and 0.7) for minority senses (0–20%), but excellent performance (F1-scores above 0.9) for majority senses (80–100%). Table 5 also provides more details about the impact of the additional independent training data. The biggest improvements are observed for low-frequency and medium-frequency senses and for language pairs with small (<1M sentences) training sets in the **Small** condition.

4.2. Segmentation effects

Most current NMT models use some word segmentation scheme for open-vocabulary translation, which keeps frequent word forms intact but splits rare word forms into "subwords". It could be argued that word segmentation – on the source or on the target side – has an adverse effect on sense disambiguation, as the model needs to compose the meaning of a word from its subwords. Figure 2 visualizes the effects of word segmentation on sense disambiguation. Note that we report precision values here, since the target segmentation is unknown when no relevant target word can be found, making it impossible to compute recall.

Except for one language pair (EN–FR), segmenting the source word does not have any effect on translation; on average, source word splitting even slightly increases precision. On the target side, results are less clear-cut, but show the same overall tendency. Therefore, word segmentation does not seem to affect the disambiguation capabilities of an MT model.

4.3. Corpus effects

When building the training and test sets, we have split the data sources into "clean" and "noisy" ones and strived to use as many examples as possible of the former. Figure 3 (left) shows indeed that test sentences from noisy sources

Lang. pair	Precision	Weighted Prec.
CS–EN	0.8044	0.8414
DE–EN	0.7831	0.8389
FI–EN	0.8125	0.8441
FR–EN	0.7668	**0.8294**
RU–EN	0.7894	0.8386
EN–CS	0.7859	0.8394
EN–DE	0.7829	**0.8582**
EN–FI	0.8003	**0.8747**
EN–FR	0.7303	**0.8312**
EN–RU	0.7965	**0.8684**

Table 6: Weighted and unweighted precision values obtained with the **Big** models. Differences of at least 0.06 absolute are highlighted in bold.

are much harder to translate correctly than those from clean ones, with both precision and recall being affected likewise. The negative impact of noisy data sources may just be due to a disproportionately high amount of rare senses for which few examples were available in the first place. The noisy subset indeed generally consists of higher proportions of rare senses than the clean subset, but there does not seem to be a direct interaction between the two phenomena. For instance, although FI–EN, EN–FI and RU–EN have much more balanced distributions of rare senses than the other language pairs, they show similar patterns in Figure 3.

4.4. Effects of sense distinctiveness

Another factor that may influence disambiguation performance is the semantic similarity between the correct and incorrect senses: if it is hard for humans (and for specially developed models like sense embeddings) to distinguish between two senses, it might also be hard to do so for an NMT model. We test this hypothesis by splitting the test sentences in two bins, depending on whether the distance between the correct and all incorrect senses is lower or higher than the average distance value between senses observed for that language pair.

Figure 3 (right) shows that the expected effect holds for some language pairs (typically with English as target language), whereas the contrary effect is observed for other pairs (typically with English as source language). Experiments with a larger number of bins have shown similarly inconclusive outcomes.

This outcome could hint at shortcomings of the sense embeddings used in this work (Mancini et al., 2017). In future work, we plan to evaluate more recent sense embedding approaches, for instance an approach based on big pre-trained language models like BERT (Devlin et al., 2019; Scarlini et al., 2020).

Another way to assess sense distinctiveness is by including it directly in the evaluation metric, as proposed with the weighted precision score in Section 3. A comparison between the standard and the weighted precision scores is shown in Table 6. The weighted precision scores are generally higher, and the difference is proportional to the average similarity between senses (see Table 3).

5. Conclusion

In this paper, we presented an extended version of MuCoW, an automatically built evaluation benchmark for measuring WSD capabilities of machine translation systems, available for 10 language pairs. The construction of the benchmark was performed by exploiting large parallel corpora, the multilingual TurkuNLP neural parsing pipeline and the multilingual dictionary BabelNet with its language-independent sense embeddings. We provide training and test sets with known sense distributions, together with baseline scores from a recent NMT system.

We find that rare senses are still an open challenge also for the state-of-the-art Transformer model, but that adding more training data, not necessarily containing the ambiguous words of interest, may mitigate this problem. We show that word segmentation does not affect the disambiguation ability much, whereas the performance drops consistently across languages when evaluating on sentences from noisy sources.

In future work, we plan to further refine the evaluation benchmark by using more recent sense embeddings (Scarlini et al., 2020). We also plan to extend the analysis and the evaluation by including character-based NMT models and additional language pairs.

6. Acknowledgements

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, and the CSC – IT Center for Science, Finland, for computational resources. Finally, We would also like to acknowledge NVIDIA and their GPU grant.

7. Bibliographical References

Delli Bovi, C., Telesca, L., and Navigli, R. (2015). Large-scale information extraction from textual definitions through deep syntactic and semantic analysis. *Transactions of the Association for Computational Linguistics*, 3:529–543.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Flekova, L. and Gurevych, I. (2016). Supersense embeddings: A unified model for supersense interpretation, prediction, and utilization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2029–2041, Berlin, Germany. Association for Computational Linguistics.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Kumar, S., Jat, S., Saxena, K., and Talukdar, P. (2019). Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy, July. Association for Computational Linguistics.

Liu, F., Lu, H., and Neubig, G. (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.

Loureiro, D. and Jorge, A. (2019). Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Mancini, M., Camacho-Collados, J., Iacobacci, I., and Navigli, R. (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.

Navigli, R. (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.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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. Association for Computational Linguistics.

Pilehvar, M. T., Camacho-Collados, J., Navigli, R., and Collier, N. (2017). Towards a seamless integration of word senses into downstream NLP applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1857–1869, Vancouver, Canada. Association for Computational Linguistics.

Popović, M. and Castilho, S. (2019). Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII Volume 3: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Pu, X., Pappas, N., Henderson, J., and Popescu-Belis, A. (2018). Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Raganato, A., Delli Bovi, C., and Navigli, R. (2017). Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167, Copenhagen, Denmark. Association for Computational Linguistics.

Scarlini, B., Pasini, T., and Navigli, R. (2020). SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Sennrich, R., Haddow, B., and Birch, A. (2016). 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.

Shimura, K., Li, J., and Fukumoto, F. (2019). Text categorization by learning predominant sense of words as auxiliary task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1109–1119, Florence, Italy, July. Association for Computational Linguistics.

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., and Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163:955–971.

Tang, G., Sennrich, R., and Nivre, J. (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, Brussels, Belgium. Association for Computational Linguistics.

Tang, G., Sennrich, R., and Nivre, J. (2019). Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435, Hong Kong, China. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Zhong, Z. and Ng, H. T. (2012). Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–282, Jeju Island, Korea. Association for Computational Linguistics.

8. Language Resource References

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceed-*

ings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation, Heidelberg, Germany*.

Eisele, A. and Chen, Y. (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).

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

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Rios, A., Mascarell, L., and Sennrich, R. (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.

Rios, A., Müller, M., and Sennrich, R. (2018). The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Tiedemann, J. (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).