The MuCoW word sense disambiguation test suite at WMT 2020

Yves Scherrer **Alessandro Raganato** **Jörg Tiedemann**
University of Helsinki
{name.surname}@helsinki.fi

Abstract

This paper reports on our participation with the MuCoW test suite at the WMT 2020 news translation task. We introduced MuCoW at WMT 2019 to measure the ability of MT systems to perform word sense disambiguation (WSD), i.e., to translate an ambiguous word with its correct sense. MuCoW is created automatically using existing resources, and the evaluation process is also entirely automated. We evaluate all participating systems of the language pairs English → Czech, English ↔ German, and English → Russian and compare the results with those obtained at WMT 2019. While current NMT systems are fairly good at handling ambiguous source words, we could not identify any substantial progress – at least to the extent that it is measurable by the MuCoW method – in that area over the last year.

1 Introduction

At WMT 2019, we introduced the MuCoW (*multilingual contrastive word sense disambiguation*) test suite (Raganato et al., 2019) and evaluated the news task submissions of nine translation directions with it.[1] We observed that systems generally performed quite well on word sense disambiguation, but found a big gap between in-domain and out-of-domain disambiguation performance for some translation directions, in particular with constrained systems.

For WMT 2020, we reuse the same test suite for the same language pairs. This gives us the opportunity to measure the advancement of machine translation within a year. We expect the larger training data sets and the model improvements to have a small but positive impact on translation quality in general, and word sense disambiguation performance in particular.

[1]The MuCoW test suite is available at http://github.com/Helsinki-NLP/MuCoW.

2 The MuCoW test suite

MuCoW (Raganato et al., 2019) is a language-independent method for automatically building test suites to assess the capabilities of MT systems to disambiguate between ambiguous words in the source language. The version of MuCoW used for WMT 2019 involves the following steps:

1. Identify ambiguous source nouns and their translations, using word-aligned and tagged parallel corpora from the OPUS collection (Tiedemann, 2012).

2. Cluster the translations into senses. First, we query BabelNet (Navigli and Ponzetto, 2012), a wide-coverage multilingual encyclopedic dictionary, to assign senses (synsets) to words. Second, we refine the results with the SW2V sense embeddings (Mancini et al., 2017).

3. Select sentences with ambiguous words and assign them sets of correct and incorrect target translations.

We evaluated the systems participating in the WMT 2019 news translation task with MuCoW for the language pairs English → Czech, English ↔ German, English ↔ Finnish, English ↔ Russian, and English ↔ Lithuanian.

A substantial amount of MuCoW sentences and senses come from the OpenSubtitles2018 corpus, but most systems participating at WMT are tuned towards the news domain and therefore are not expected to handle lexical choices of colloquial speech reliably. Therefore, 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 movie subtitles.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 365–370
Online, November 19–20, 2020.

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 1: Examples of test suite instances of the English–German 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	Source words	Target synsets	In-dom synsets	Out-dom synsets	Sentences
EN–CS	98	200	29	171	1843
EN–DE	176	362	220	142	3337
DE–EN	217	461	329	132	4268
EN–RU	97	199	40	163	1814

Table 2: Sizes of the MuCoW data sets compiled for WMT 2019 and 2020.

In Raganato et al. (2020), we report on an extended version of MuCoW that covers the following aspects:

- The selection of data sources is improved 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.
- In addition to test sets, the composition of training data is also defined to guarantee that competing translation models are evaluated on fair grounds.

Since it was not possible to restrict the training data of participating WMT systems, we decided to reuse the WMT 2019 version again for WMT 2020, with exactly the same sentences. This allows us to trace the year-over-year evolution of translation quality with respect to lexical disambiguation. Therefore, the MuCoW analysis is restricted to the language pairs and translation directions that were already part of the WMT news task in 2019, namely English → Czech, English ↔ German, and English → Russian.

MuCoW data sets are created specifically for each language pair and translation direction (for details, see Raganato et al., 2019). Each entry consists of a sentence in the source language, the ambiguous source word, a list of correct target words (the correct target synset), a list of incorrect target words (the incorrect target synset), and information about the domain of the synsets. The participants only see the source sentences, not the metadata. Table 1 shows a few example sentences taken from the English–German test suite. The main statistics of the test suites used for WMT 2020 are reported in Table 2.

3 Evaluation and Results

The source language sentences were sent to the WMT participants as part of the test set, and we received the translations in the target language for evaluation. We then checked if any of the correct or incorrect target words listed in the metadata file could be identified in the translation output.

Although the sentences were selected to contain the uninflected base forms both in the source and target languages, we could not assume that all translation systems would output base forms. Hence, if neither correct nor incorrect target words could be identified in the tokenized translations, we lemmatized them and searched the target words again in the lemmatized version.[2] Depending on the morphological properties of the target language, lemmatization substantially increased the coverage (see Table 3). Between 2019 and 2020, the average coverage has remained constant

[2]We used the Turku neural lemmatizer with pretrained models (Kanerva et al., 2019).

Language pair	Avg. coverage (tokenized)	Avg. coverage (tok. + lemmatized)
EN–CS	63.16%	75.82%
	61.77%	*74.87%*
EN–DE	69.43%	72.08%
	66.52%	*69.26%*
DE–EN	83.10%	84.41%
	83.06%	*84.51%*
EN–RU	65.13%	80.13%
	58.88%	*73.29%*

Table 3: Average coverage of target words among WMT 2019 (in gray italics) and WMT 2020 (in black) primary submissions.

for DE–EN, slightly increased for EN–CS and EN–DE, and substantially increased for EN–RU. We assume that these increases are mostly due to the different number and composition of the submissions.

We report precision, recall and F1-score for in-domain senses and out-of-domain senses separately. Precision and recall are computed as follows:[3]

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

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

The results are shown in Tables 4 to 7, with WMT 2019 and 2020 submissions side-by-side.

For all four examined translation directions, the best 2019 results were beaten in 2020. However, one of the best-performing systems in 2019, *Facebook_FAIR*, did not participate in 2020. The *Facebook_FAIR* system is characterized by high precision rates, whereas the winning 2020 systems (such as *Tohoku-AIP-NTT* or *Online-G*) benefit from higher recall. This shift suggests that the denominator of the precision computation comes closer to the one of the recall computation, or in other words that the translations themselves become more accurate. Further analysis will be required to substantiate this claim.

Interesting year-over-year comparisons can be observed for the *Online-G* system: it produces almost identical results in both years for English–German and English–Russian, but shows substantial improvements for the German–English direction.

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

The overall result distributions show a slight upward trend in WSD performance for English–German and German–English, but less so for English–Czech and English–Russian. Since the participating systems differed over the years, it is of course difficult to draw any reliable conclusions.

For most language pairs, the in-domain and out-of-domain synsets produce similar rankings. Just like in 2019, English–Czech is an exception, where – contrarily to all expectations – an online system shows the best in-domain performance and a research system the best out-of-domain performance.

4 Conclusion

In this paper, we report our participation with the MuCoW test suite at the WMT 2020 news translation task. MuCoW is an automatically built WSD test suite for machine translation that relies on large parallel corpora, the multilingual lexical resource BabelNet and language-independent synset embeddings.

We find that state-of-the-art NMT systems are fairly good at handling ambiguous source words, but that no substantial progress – at least to the extent that it is measurable by the MuCoW method – has been made in that area over the last year. Among the top-performing systems, we observe a shift from high precision to high recall, hinting at general improvements in translation quality. It will therefore 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).

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 CSC – IT Center for Science, Finland, for computational resources.

English–Czech	In-domain synsets			Out-of-domain synsets			All synsets		
Submission	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
SRPOL	97.15	84.45	90.36	**80.38**	**73.78**	**76.94**	**83.22**	**75.67**	**79.27**
CUNI-Transformer	95.53	84.23	89.52	80.00	72.75	76.21	82.65	74.76	78.51
CUNI-T2T-2018	96.80	85.82	90.98	79.54	71.78	75.46	82.55	74.26	78.19
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-DocTransformer	97.19	85.51	90.98	79.06	71.08	74.86	82.23	73.65	77.70
eTranslation	95.20	85.61	90.15	76.13	70.15	73.02	79.48	72.92	76.06
OPPO	96.03	86.43	90.98	74.35	68.55	71.33	78.23	71.81	74.88
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*
UEDIN-CUNI	95.98	85.36	90.36	71.24	66.07	68.56	75.69	69.65	72.54
Online-A	95.49	83.51	89.10	69.89	67.28	68.56	74.34	70.33	72.28
Online-G	96.77	85.11	90.57	68.74	65.41	67.04	73.76	69.17	71.39
Online-Y	*97.57*	*84.86*	*90.77*	*61.57*	*63.73*	*62.63*	*67.93*	*68.03*	*67.98*
Online-Z	97.57	84.86	90.77	61.67	61.01	61.34	68.19	65.82	66.98
parfda	*95.02*	*75.27*	*84.00*	*68.16*	*58.44*	*62.93*	*72.85*	*61.57*	*66.74*
Online-B	**98.44**	**88.11**	**92.99**	57.50	59.80	58.63	65.12	65.74	65.43
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*
zlabs-nlp	95.55	84.59	89.73	47.21	47.68	47.45	56.61	55.65	56.13

Table 4: Results for English–Czech. WMT 2019 submissions are displayed in gray italics.

English–German	In-domain synsets			Out-of-domain synsets			All synsets		
Submission	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Tohoku-AIP-NTT	83.17	**77.09**	80.01	55.53	**57.93**	**56.71**	73.82	**71.11**	**72.44**
Facebook_FAIR	***83.43***	*76.99*	***80.08***	***56.29***	*55.10*	*55.69*	***74.48***	*70.05*	*72.19*
Online-B	82.52	77.27	79.81	52.48	56.45	54.39	72.40	70.88	71.63
Microsoft-sentence-level	*83.18*	*77.14*	*80.05*	*52.81*	*51.92*	*52.36*	*73.31*	*69.27*	*71.23*
OPPO	81.81	76.48	79.05	52.58	55.23	53.87	72.01	69.89	70.93
Huoshan_Translate	82.05	77.16	79.53	50.24	53.32	51.73	71.50	69.89	70.68
eTranslation	81.99	75.36	78.53	51.44	52.77	52.09	71.82	68.38	70.05
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*
AFRL	81.82	73.96	77.69	45.73	45.33	45.53	70.16	65.28	67.63
Online-G	81.44	73.76	77.41	46.61	45.44	46.02	70.21	65.09	67.55
Online-G	*81.44*	*73.76*	*77.41*	*46.61*	*45.44*	*46.02*	*70.21*	*65.09*	*67.55*
Online-A	81.26	73.45	77.16	45.72	43.05	44.35	70.00	64.09	66.92
DFKI-NMT	*80.70*	*74.37*	*77.41*	*44.95*	*42.04*	*43.44*	*69.54*	*64.39*	*66.87*
PROMT_NMT	79.62	72.84	76.08	42.65	47.05	44.74	67.24	65.24	66.23
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*
UEDIN	78.55	75.47	76.98	37.42	39.56	38.46	65.61	64.90	65.25
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*
Online-Z	75.61	69.71	72.54	41.06	43.03	42.02	64.18	61.62	62.87
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*
zlabs-nlp	77.33	66.55	71.54	36.78	28.87	32.35	65.36	54.70	59.55
TartuNLP-c	*77.32*	*66.29*	*71.38*	*33.02*	*26.13*	*29.17*	*64.34*	*53.85*	*58.63*
WMTBiomedBaseline	73.59	57.02	64.25	31.91	15.52	20.88	63.33	42.82	51.09
EN_DE_Task	*64.54*	*23.14*	*34.06*	*38.41*	*5.64*	*9.84*	*59.43*	*16.62*	*25.97*

Table 5: Results for English–German. WMT 2019 submissions are displayed in gray italics.

German–English	In-domain synsets			Out-of-domain synsets			All synsets		
Submission	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Online-G	80.35	**86.75**	**83.43**	51.37	**75.37**	**61.10**	72.78	**84.40**	**78.16**
Facebook_FAIR	***80.78***	*85.80*	*83.21*	***52.77***	*72.56*	***61.10***	***73.55***	*82.99*	*77.99*
Tohoku-AIP-NTT	80.52	86.32	83.32	48.56	72.84	58.27	72.21	83.62	77.50
OPPO	80.03	86.14	82.97	47.83	71.74	57.39	71.69	83.25	77.04
Online-B	80.36	83.75	82.02	48.79	69.68	57.39	72.16	80.88	76.27
Huoshan_Translate	78.11	86.00	81.86	45.05	71.06	55.14	69.53	83.06	75.70
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-A	77.86	83.58	80.62	41.39	64.50	50.42	68.50	79.91	73.77
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*
PROMT_NMT	75.14	83.75	79.21	38.74	60.85	47.34	65.95	79.33	72.02
Online-A	*75.77*	*83.08*	*79.26*	*37.47*	*63.15*	*47.04*	*65.87*	*79.40*	*72.00*
UEDIN	75.57	85.08	80.05	32.86	57.69	41.87	64.84	80.23	71.72
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*
Online-Z	73.89	80.53	77.07	38.32	63.67	47.85	64.56	77.34	70.37
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*
WMTBiomedBaseline	69.23	70.34	69.78	23.05	22.63	22.84	59.54	60.05	59.79
zlabs-nlp	62.87	76.50	69.02	19.67	30.10	23.79	52.87	67.53	59.30

Table 6: Results for German–English. WMT 2019 submissions are displayed in gray italics.

English–Russian	In-domain synsets			Out-of-domain synsets			All synsets		
Submission	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Online-G	**96.11**	89.64	92.76	**75.44**	74.52	**74.98**	**80.46**	78.35	**79.39**
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	94.97	89.01	91.89	63.86	71.67	67.54	71.35	76.44	73.81
OPPO	95.07	90.84	92.90	62.31	69.38	65.65	70.42	75.33	72.79
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-A	94.78	90.55	92.62	58.24	69.21	63.25	67.18	75.34	71.03
Ariel197197	95.66	85.97	90.56	61.40	66.77	63.97	69.70	72.12	70.89
Online-Y	*95.37*	*91.38*	***93.33***	*57.47*	*69.02*	*62.72*	*66.80*	*75.51*	*70.89*
PROMT_NMT	94.25	90.77	92.47	60.61	65.69	63.05	69.15	72.63	70.84
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*
Online-Z	95.80	88.83	92.18	53.95	60.97	57.24	64.56	69.13	66.76
zlabs-nlp	94.99	89.27	92.04	51.56	60.78	55.79	62.54	69.27	65.73
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 7: Results for English–Russian. WMT 2019 submissions are displayed in gray italics.

References

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.

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*.

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.

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.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 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.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3668–3675, Marseille, France. European Language Resources Association.

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.

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).