The Word Sense Disambiguation Test Suite at WMT18

Annette Rios[1] **Mathias Müller**[1] **Rico Sennrich**[1,2]

[1]Institute of Computational Linguistics, University of Zurich
{rios,mmueller}@cl.uzh.ch

[2]School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk

Abstract

We present a task to measure an MT system's capability to translate ambiguous words with their correct sense according to the given context. The task is based on the German–English Word Sense Disambiguation (WSD) test set ContraWSD (Rios Gonzales et al., 2017), but it has been filtered to reduce noise, and the evaluation has been adapted to assess MT output directly rather than scoring existing translations. We evaluate all German–English submissions to the WMT'18 shared translation task, plus a number of submissions from previous years, and find that performance on the task has markedly improved compared to the 2016 WMT submissions (81%→93% accuracy on the WSD task). We also find that the unsupervised submissions to the task have a low WSD capability, and predominantly translate ambiguous source words with the same sense.

1 Introduction

Ambiguous words are often difficult to translate automatically, since the MT system has to decide which sense is correct in the given context. Errors in lexical choice can result in bad or even incomprehensible translations. However, document-level metrics, such as BLEU (Papineni et al., 2002) are not fine-grained enough to assess this type of error.

Early evaluations have shown that neural machine translation (NMT) produces translations that are substantially more *fluent*, i.e. more grammatical and natural, than the previously dominant phrase-based/syntax-based statistical models, but results are more mixed when comparing *adequacy*, the semantic faithfulness of the translation to the original (Bojar et al., 2016; Bentivogli et al., 2016; Castilho et al., 2017; Klubička et al., 2017).

For example, in the fine-grained human evaluation by Klubička et al. (2017), *mistranslations* were the most frequent error category for the NMT system they evaluated, whereas fluency errors dominated in phrase-based machine translation.[1] Our aim is to quantify one aspect of adequacy, word sense disambiguation (WSD), in a reproducible and semi-automatic way, to track progress over time and compare different types of systems in this respect.

We present a German→English test set to semi-automatically assess an MT systems performance on word sense disambiguation. The test set is based on ContraWSD (Rios Gonzales et al., 2017), but has been further filtered to reduce noise, and we use a different evaluation protocol. Instead of scoring a set of translations and measuring whether the reference translation is scored highest, we base the evaluation on the 1-best translation output to make the evaluation applicable to black-box systems. We report results on all German→English submissions to the WMT 2018 shared translation task (Bojar et al., 2018), plus a number of baseline systems from previous years.

2 Test Suite

Rather than measuring word sense disambiguation against a manually defined sense inventory such as those in Wordnet (Miller, 1995), we perform a task-based evaluation, focusing on homonyms whose different senses have distinct translations.[2]

[1]Note that, while mistranslations were the most frequent error category in NMT, their absolute number was still lower in the NMT system output than in the phrase-based one.

[2]Other task-based evaluation sets for word sense disambiguation include (Lefever and Hoste, 2013; Gorman et al., 2018).

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 594–602

https://doi.org/10.18653/v1/W18-6437

The collection of test cases consists of 3249 German–English sentence pairs where the German source contains one of 20 ambiguous words that have more than one possible translation in English.[3] We have associated the 20 ambiguous words with a total of 45 word senses, and extracted up to 100 examples for each sense.

The set of ambiguous words and sentence pairs are based on the test set described in (Rios Gonzales et al., 2017).[4] The original test set was designed to use scoring for the evaluation, however, in the present task we let the systems translate the source sentences, and evaluate the translation output. This change in evaluation protocol required further filtering of the original test set, specifically, the removal of German words with an English translation that covers multiple senses. For instance, the original test set contains *Stelle* with two English senses: *job* and *place*. Both meanings can be translated as *position*, in which case we would not be able to assess the translation as correct or wrong, therefore *Stelle* was removed from our set of ambiguous words.

Since for most ambiguous words, one or more of their meanings are relatively rare, a large amount of parallel text is necessary to extract a sufficiently balanced number of examples.[5] The correct translation is automatically determined for each pair through the reference translation. Table 1 lists all the ambiguous German words in the test set with with their translations in English. We base our statistics on the number of ambiguous source words, which is slightly higher (3363) than the number of sentences (3249). Sentence pairs where the reference translation contains more than one possible sense as a translation have been removed. For instance, if a given reference contains the word *investment* as a translation for *Anlage*, but also *attachment* as a translation of another source word, this sentence pair cannot be part of the test set, since word alignment would be required to assess it correctly.

The evaluation is semi-automatic: We automatically check for each sentence in the MT output if one of the correct translations of the ambiguous word is present, and if the output contains one of the other possible translations of the word, i.e. if it has been translated with one of its other senses. Note that we check for more variation in the automatic matching than shown in Table 1, e.g. for *Absatz - sales*, we also consider verbal forms such as *sold, sells, selling etc.* as correct, using manually created lists of valid translations.[6]

There are four possible outcomes of this automatic evaluation:

1. we find only instances of the correct translations → counts as correct[7]
2. we find only instances of the other translations → counts as wrong
3. we find both the correct and one of the other translations → manual inspection
4. we find none of the known translations → manual inspection

2.1 Manual Evaluation Protocol

The large majority of translation outputs could be categorized as correct or wrong automatically. For the remaining approximately 5%, we manually assigned a label. Overall, around 25% of these were labelled as correct.

Case 3 typically indicates that the same ambiguous source word occurs multiple times in the input, and a manual annotator provided the number

[3] The test set and evaluation scripts are available from https://github.com/a-rios/ContraWSD/tree/master/testsuite_wmt18

[4] The identification of ambiguous words and senses was performed with the help of lexical translation probabilities.

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

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

[6] Since we do not use word alignment, there is a risk that we mistakenly match a translation from another part of the sentence. However, this is only a problem in the rare case where, at the same time, the ambiguous source word itself is not translated into a known translation, since conflicting matches trigger a manual inspection.

[7] If there are multiple instances of the ambiguous source word in the sentence, we automatically count the number of correct translations to assign credit.

word	translations/senses			
	sense 1	sense 2	sense 3	sense 4
Absatz	heel	paragraph	sales	
Anlage	attachment, annex	installation, facility, plant	investment	
Annahme	acceptance, approval	assumption, conjecture		
Aufgabe	abandonment, surrender	task, exercise		
Auflösung	dissolution, liquidation	resolution		
Decke	blanket, cover	ceiling		
Einsatz	bet	commitment	usage, application	
Gericht	court, tribunal	dish, meal		
Himmel	heaven	sky		
Karte	card	menu	ticket	map
Kurs	course, class	price, rate		
Lager	storage, stock	camp		
Opfer	victim	sacrifice		
Preis	prize	price, cost, fee		
Rat	advice, counsel	council, board		
Raum	region,area	room, space		
Schlange	serpent, snake	queue, line		
Ton	tone, sound	clay		
Tor	door, gate	goal		
Wahl	election	choice, selection		

Table 1: List of ambiguous German words, and the English translations of their different senses, included in the test suite.

source	Im Allgemeinen lässt sich deshalb mit Recht behaupten, dass – mit der richtigen Beratung und Sorgfalt – Hedge-Fund-**Anlagen** nicht zwangsläufig risikoreicher sind als traditionelle **Anlagen**.
reference	It is therefore fair to say that properly advised hedge fund **investments** are, generally speaking, not necessarily riskier than traditional **investments**.
MT translation	In general, therefore, it is fair to say that, with the right advice and care, hedge fund **assets** are not necessarily more risky than traditional **plants**.

Table 2: Example sentence pair for ambiguous word *Anlagen* with translation from uedin-nmt-2017. The first translation *assets* is correct, the second (*plants*) wrong.

of correct translations. See Table 2 with an example from one of the baseline systems, where the ambiguous word *Anlage* occurs twice, both times in the financial sense. The MT system translates the first form correctly, but the second with one of its other meanings, *plant*.

Case 4 can indicate that the ambiguous source word was translated into a variant not covered by our automatic patterns, or left untranslated.[8] Manual assessment by the main author is used to distinguish between the two.

3 Evaluation

We present results for all submissions to the WMT'18 shared translation task for German→English. In addition, we include several baseline systems in our evaluation to track performance over time. We report results for Edinburgh's WMT'16 and WMT'17 submitted neural systems for German→English (Sennrich et al., 2016, 2017), which were ranked first in 2016, and tied first in 2017.[9] We also include Edinburgh's WMT'16 syntax-based system (Williams et al., 2016), ranked tied second in 2016, to compare the now dominant neural systems to a more traditional SMT system.

We report the WSD accuracy for each system, in two variants: automatic and full. For automatic accuracy only case 1 is considered correct, and cases 2–4 are considered wrong. Full accuracy considers some cases 3 and 4 (where both a correct and an incorrect translation, or none of the listed translations, are found) correct, if they were found to be correct upon manual inspection. We also report BLEU scores on newstest2018, and on the WSD test suite, for comparison.

4 Results

Results on the WSD test suite are shown in Table 3. Table 4 shows an error analysis with two categories, distinguishing between predicting the wrong sense, and leaving the ambiguous source word untranslated. Globally, we observe a strong correlation between WSD accuracy and BLEU on the WSD test suite (Kendall's $\tau = 0.91$), and a smaller (but still strong) correlation between WSD accuracy and BLEU on newstest2018 ($\tau = 0.72$). However, there are some notable differences between BLEU and WSD accuracy. Especially some unconstrained, anonymous systems (online-A/B/G/Y) perform better on the WSD test suite than newstest2018 relative to other systems, which is likely due to differences in domain focus and training data: most constrained systems built for the shared task use monolingual news data for domain adaptation, whereas the online systems likely do not. At the same time, the online systems may be using extra training resources, and we cannot rule out that they train on corpora from which the WSD test suite is extracted.

The unsupervised systems RWTH-UNSUPER and LMU-unsup, as well as the anonymous rule-based system online-F clearly fall behind. In many cases, these systems stick to one translation of a given ambiguous word. This becomes obvious when looking at the number of cases where the translation contains one of the other meanings of the translated words. The less common a given sense, the more likely it is translated with one of its other meanings - this is true for all systems, but more pronounced in the unsupervised models. Not only do they translate words with a wrong meaning more often, they seem to have learned some spurious correlations. For instance, the German word *Preis* (*price/prize*) was translated in almost all cases as *call* by LMU-unsup. Generally, the unsupervised systems tend to translate words in a deterministic fashion, i.e. they use mostly the same translation for an ambiguous source word, regardless of context.

We observe that there is little difference in WSD accuracy between the syntax-based and neural uedin systems from 2016, even though the neural system achieves a substantially higher BLEU score. This is consistent with human comparisons of statistical and neural systems at the time, which found large improvements in fluency, but only small differences in adequacy, or specifically the number of mistranslations (Bojar et al., 2016; Castilho et al., 2017; Klubička et al., 2017). Interestingly, we observe major improvements in lexical choice since the 2016 systems, with a jump of 5 percentage points in 2017, and another 8 percentage points by the best system in 2018.

While these experiments were not under controlled data conditions[10], we believe that this im-

[8] This includes cases where the original source word is used in the translation.

[9] Available at http://data.statmt.org/wmt16_systems/ and http://data.statmt.org/wmt17_systems/

[10] There was a moderate improvement in the amount of training data in 2017 through the inclusion of the Rapid corpus of EU press releases (+20%), and a large increase in 2018

system	WSD accuracy		BLEU	
	automatic	full	newstest2018	WSD test suite
uedin-syntax-2016	79.7	81.3	36.1	26.9
uedin-nmt-2016	79.8	81.1	41.3	27.7
uedin-nmt-2017	84.9	86.3	43.5	30.5
RWTH	92.4	93.6	48.4	33.6
UCAM	91.1	92.4	48.0	32.9
online-B	89.4	91.3	43.9	32.5
NTT	89.7	91.2	46.8	32.6
JHU	88.9	90.3	45.3	31.7
online-Y	88.0	89.8	39.5	30.9
MLLP-UPV	88.4	89.7	45.1	30.7
uedin	87.1	88.8	43.9	30.8
Ubiqus-NMT	86.7	88.3	44.1	31.0
online-A	86.6	88.0	40.6	29.7
online-G	85.4	86.9	31.9	29.1
NJUNMT-private	84.3	86.0	38.3	28.2
LMU-nmt	80.4	81.7	40.9	28.1
online-F	50.7	51.4	22.0	15.8
RWTH-UNSUPER	44.9	47.2	18.6	11.4
LMU-unsup	42.6	43.3	17.9	10.0

Table 3: Results on WSD test suite. WSD accuracy before and after manual inspection, and BLEU on newstest2018, and on references from WSD test suite.

system	wrong sense	untranslated
uedin-syntax-2016	17.4	1.3
uedin-nmt-2016	16.5	2.4
uedin-nmt-2017	11.7	2.1
RWTH	5.2	1.2
UCAM	6.4	1.2
online-B	6.5	2.1
NTT	7.0	1.8
JHU	8.5	1.2
online-Y	9.0	1.2
MLLP-UPV	9.5	0.8
uedin	10.1	1.2
Ubiqus-NMT	9.3	2.3
online-A	11.0	1.0
online-G	11.6	1.5
NJUNMT-private	9.3	4.7
LMU-nmt	16.3	2.1
online-F	47.8	0.7
RWTH-UNSUPER	48.9	3.9
LMU-unsup	49.8	6.9

Table 4: Proportion of ambiguous words translated with the wrong sense, or left untranslated (in %).

provement is only partially explainable by the increase in the amount of training data. We highlight a number of systems to illustrate this point.

Paracrawl is a noisy resource, and most submission systems report using a filtered version of it. Ubiqus-NMT does not use Paracrawl at all, and is thus comparable to uedin-nmt-2017 in terms of training data, but outperforms it in WSD accuracy. This is even more impressive considering that Ubiqus-NMT is based on a single model, outperforming the reranked ensembles of uedin-nmt-2017.

A second interesting comparison is that between different architectures. LMU-nmt is based on a shallow RNN encoder-decoder, similar to uedin-nmt-2016, and exhibits a similarly low WSD accuracy. Most submissions are based on deep Transformer or RNN architectures, and show a higher WSD accuracy. Neural network depth was also one of the main differences between uedin-nmt-2016 and uedin-nmt-2017, and our results indicate that this is an important factor for lexical choice. Experiments by Tang et al. (2018), conducted in parallel to this work, on WMT17 training data also show that neural architectures

through the inclusion of Paracrawl (+700%).

play an important role in the performance on WSD, with a substantial lead for the Transformer over the tested RNN and CNN architectures.

The error analysis in Table 4 exposes other differences between systems. The rule-based system online-F is least prone to leaving the ambiguous source words untranslated (0.7%), while this is a more serious problems in the unsupervised systems (up to 6.9%) and some neural systems (up to 4.7%). It has been argued that SMT, which uses a coverage mechanism during decoding, is less prone to undertranslation than NMT (Tu et al., 2016). On the WSD test set, we find that uedin-nmt-2016 leaves more of the ambiguous words untranslated (2.4%) than the contemporaneous uedin-syntax-2016 (1.3%), but most NMT systems submitted to this year's shared translation task improve upon this number. While this is a very narrow evaluation of the undertranslation problem (only on one data set, and looking at specific source words), we consider it encouraging that we could measure some progress.

5 Conclusions

We present a targeted evaluation of 16 systems regarding their performance in lexical choice. A comparison against a baseline consisting of the top ranked systems from WMT 2016 and 2017 for German-English shows that translation models in general have improved substantially. Furthermore, we observe that unsupervised systems are at a clear disadvantage when it comes to word sense disambiguation: they are less flexible and tend to stick to one translation of a given ambiguous word, regardless of context.

The current study is focused on a small set of 20 ambiguous nouns and 45 word senses, and a large-scale test set is created by extracting 3249 sentence pairs containing one of these word senses from various parallel corpora. This focus on ambiguous source words without lexical overlap between word senses in the target language allowed us to define an evaluation protocol that is mostly automatic: manual inspection was only necessary for about $\approx 5\%$ of sentences, and had little effect on the ranking. However, this narrow focus also comes with limitations, and it would be interesting to evaluate word sense disambiguation on a larger set of words, and including other parts-of-speech such as verbs and adverbs, which constituted a substantial proportion of lexical choice errors in previous analyses of MT systems (Williams et al., 2015).

Acknowledgement

We are grateful to the Swiss National Science Foundation (SNF) for support of the project CoNTra (grant number 105212_169888). We thank the anonymous reviewers for their helpful feedback.

References

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

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

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A Comparative Quality Evaluation of PBSMT and NMT using Professional Translators. In *Proceedings of Machine Translation Summit XVI*, Nagoya, Japan.

Kyle Gorman, Gleb Mazovetskiy, and Vitaly Nikolaev. 2018. Improving homograph disambiguation with supervised machine learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Filip Klubička, Antonio Toral, and Víctor M. Sánchez-Cartagena. 2017. Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation. *The Prague Bulletin of Mathematical Linguistics*, 108:121–132.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

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

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Barry Haddow, and Ondřej Bojar. 2016. Edinburgh's Statistical Machine Translation Systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, pages 399–410, Berlin, Germany. Association for Computational Linguistics.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 199–209, Lisbon, Portugal. Association for Computational Linguistics.

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

Appendix A

system	WSD accuracy	
	automatic	full
uedin-syntax-2016	77.6	79.3
uedin-nmt-2016	77.7	79.1
uedin-nmt-2017	83.0	84.6
RWTH	91.8	93.2
UCAM	90.3	91.7
online-B	88.5	90.6
NTT	88.5	90.3
JHU	87.7	89.3
online-Y	87.1	89.1
MLLP-UPV	87.2	88.7
uedin	85.6	87.5
Ubiqus-NMT	85.3	87.2
online-A	85.4	86.9
online-G	84.4	86.1
NJUNMT-private	83.1	85.1
LMU-nmt	78.1	79.6
online-F	48.3	49.0
RWTH-UNSUPER	38.5	41.2
LMU-unsup	38.3	38.9

Table 5: Results on WSD test suite, ignoring sentences from WMT dev/test data. WSD accuracy before and after manual inspection.