UAlberta at SemEval-2021 Task 2: Determining Sense Synonymy via Translations

Bradley Hauer, Hongchang Bao, Arnob Mallik, Grzegorz Kondrak
Alberta Machine Intelligence Institute, Department of Computing Science
University of Alberta, Edmonton, Canada
{bmhauer,hongchan,amallik,gkondrak}@ualberta.ca

Abstract

We describe the University of Alberta systems for the SemEval-2021 Multilingual and Cross-lingual Word-in-Context (MCL-WiC) disambiguation task. We explore the use of translation information for deciding whether two different tokens of the same word correspond to the same sense of the word. Our focus is on developing principled theoretical approaches which are grounded in linguistic phenomena, leading to more explainable models. We show that translations from multiple languages can be leveraged to improve the accuracy on the WiC task.

1 Introduction

This paper describes the University of Alberta systems for SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (Martelli et al., 2021). We focus on the monolingual (English) variant of the task, which is the same as the original WiC task (Pilehvar and Camacho-Collados, 2018). An instance of the WiC task consists of two sentences that share a *focus word* in common; the word may be inflected differently in each sentence (e.g. "they had searched his flat a few days before" and "the production of lithium from salt flats") but will share the same lemma and part of speech. A WiC task system must decide, given such a pair of sentences, whether the *focus tokens* have the same meaning in both sentences. Systems are compared in terms of their accuracy, the percentage of test instances correctly identified as TRUE (same meaning) or FALSE (different meaning). The dataset includes training, development, and testing splits; as our methods are unsupervised, we do not use the training data.

The goal of this paper is an exploration of the use of translation information for the WiC task. The intuition underlying our work is that distinctions in meaning tend to be reflected in distinctions in translation. We have previously presented methods leveraging translation information to improve word sense disambiguation (Luan et al., 2020), and most frequent sense detection (Hauer et al., 2019), and have demonstrated that word senses which share translations are, in general, semantically related (Hauer and Kondrak, 2020a). We have also presented theoretical formalizations of lexico-semantic phenomena which view synonymy and translation as two aspects of semantic equivalence (Hauer and Kondrak, 2020b). Our team additionally presented a method based on translation information (Hauer et al., 2020) for the SemEval-2020 Task 2 on Predicting Multilingual and Cross-Lingual Lexical Entailment (Glavaš et al., 2020). In this task, we investigate whether translation can be used to detect semantic equivalence in context, just as in the aforementioned prior task we investigated whether translation can be used to detect lexical entailment between word types. Our focus is on developing principled theoretical approaches which are grounded in linguistic phenomena, leading to more explainable models.

Our more complex methods depend upon a mapping between word senses and translations, as different senses of a word often translate differently. We obtain such a mapping from BabelNet (Navigli and Ponzetto, 2012), which combines information from Princeton WordNet (Fellbaum, 1998), multilingual lexical resources, and translations produced by MT models. WordNet is comprised of synonym sets, or *synsets*, which BabelNet enriches with translations. Each of the resulting multi-lingual synsets, or *multi-synsets*, contain lexicalizations of a single concept in various languages, allowing the translations of a given sense of a word to be identified. We treat BabelNet as an imperfect implementation of a universal multi-wordnet with the theoretical properties described by Hauer and Kondrak (2020b).

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 763–770
Bangkok, Thailand (online), August 5–6, 2021.

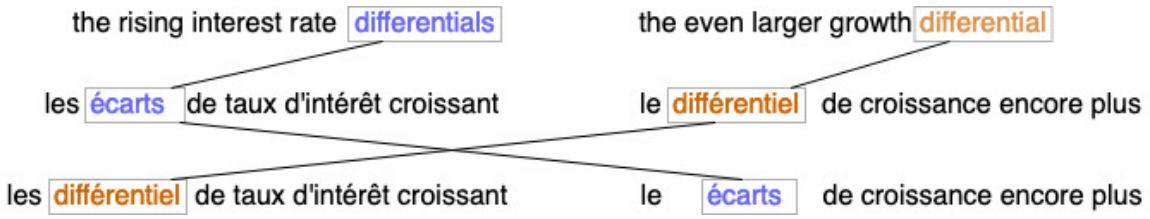

Figure 1: An example of the "translation criss-cross" described in Section 3.2.

Our results can be interpreted as a proof-of-concept for the use of contextual translations as indicators of semantic similarity. We show that the methods that we develop for the WiC task can leverage translations to improve over baselines, especially when multiple target languages are considered. While it is not our objective to compete with state-of-the-art supervised methods, we consider this to be a positive result, and a strong lead for future work on contextual semantic analysis.

This paper is structured as follows: Section 2 provides an overview of relevant prior literature. Section 3 discusses the theoretical model underlying our work. Section 4 outlines our methods. Section 5 describes our experiments and results.

2 Related Work

Methods for WiC task can be roughly divided into two paradigms: contextualized-embedding-based systems, and word sense disambiguation-based systems. Pilehvar and Camacho-Collados (2018) introduce the WiC dataset as a benchmark for evaluating context sensitive word representations. Soler et al. (2019) achieve improvements by combining similarity scores from different types of contextual word and sentence embeddings. Liu et al. (2020) propose a method to enhance contextual representations by leveraging other pre-trained contextual or static embeddings.

Another approach to WiC task is to employ a word sense disambiguation (WSD) system to tag the target words with senses from a pre-defined sense inventory and subsequently make a decision based on the predicted synsets of the target words. Loureiro and Jorge (2019b) use the LMMS sense embeddings (Loureiro and Jorge, 2019a) to disambiguate the target words. A simple approach of checking if the disambiguated senses are equal lead to competitive performance in the SemDeep-5 WiC challenge (Anke et al., 2019). SENSEMBERT (Scarlini et al., 2020a) and ARES (Scarlini et al., 2020b) embeddings, when used as features in a BERT-based model, also achieve competitive results on the WiC task.

Our methods combine elements of both paradigms. We employ contextual embeddings in our proposed translation-based methods. However, we take the embeddings of the translations of the target words instead of the target words themselves. Similarly to WSD based approaches, our methods also analyze the common synsets of the focus tokens and their translations, with the goal of identifying a probable shared synset. The most similar prior work to our approach is that of Pesutto et al. (2020) at the graded word similarity task (Armendariz et al., 2020) of SemEval 2020, who propose a translation-based approach to evaluate the contextual similarity of a pair of words. They hypothesize that leveraging similarity information from more languages would allow greater accuracy. We follow a similar intuition in our work.

3 Theoretical Solution

We first present a theoretical solution, which provides the foundation for the development of our actual methods described in Section 4. We assume that the two source sentences S_1 and S_2 in each instance of the WiC task can be translated into any natural language as sentences T_1 and T_2. Furthermore, we assume that the literal lexical translations t_1 and t_2 of the focus word s can be identified in T_1 and T_2, respectively. For example, in Figure 1, the focus word s in the English sentences S_1 and S_2 is the noun *differential*, and word alignment identifies *écart* and *différentiel* as t_1 and t_2. Note that the two translations may have the same POS and lemma, a scenario we denote as $t_1 = t_2$.

3.1 Substitution Test

Our theoretical solution is based on the notion of the linguistic *substitution test* for verifying the synonymy of senses (Hauer and Kondrak, 2020b), which takes as input two sentences *which differ only in a single word*, and returns TRUE if and

only if the two sentences have the same meaning. In other words, it decides whether the substitution of one word with another changes the meaning of the sentence. Note that this substitution test is not sufficient to decide the WiC task, as the input sentences for this task *share* a single word, rather than *differ* in a single word. The substitution test can be implemented by consulting a native speaker, or approximated by a computer program. In Section 4, we discuss an implementation based on contextual embeddings.

An example of a valid input to the substitution test would be the sentences *I work at the plant* and *I work at the factory*. For this input, the substitution test would return TRUE, since the word substitution does not change the meaning of the sentence. The sentences *I work at the plant* and *I work at the flower* would likewise constitute a valid input; however, given these sentences, the substitution test would return FALSE, since the sentences differ semantically.

3.2 Translation Criss-Cross

In order to apply the substitution test to an instance of the WiC task, we first translate the two source input sentences S_1 and S_2 into a target language, producing two target sentences T_1 and T_2. We identify the two lexical translations t_1 and t_2 of the focus word s in T_1 and T_2. Assuming that the translations are correct and literal, the senses of s in S_1 and t_1 in T_1 will be synonymous, as well as the senses of s in S_2 and t_2 in T_2. If t_1 and t_2 have the same POS but different lemmas, we can replace t_1 with t_2 in T_1 to produce a sentence T_1' which differs from T_1 in a single word. The application of the substitution test to (T_1, T_1') returns TRUE if and only if the sense of t_2 in T_1' is synonymous with the sense of s in S_1, which implies that, in addition to s and t_1, the multi-synset containing the sense of s in S_1 must also include t_2.

Using our running example in Figure 1, T_1' would be created by replacing *écarts* with *différentiel* in T_1. This produces *les différentiel de taux d'intérêt croissant*, which, while not necessarily grammatical, can still be evaluated by the substitution test to decide whether the substitution alters the semantic content of the sentence. (Or, equivalently, whether *écart* and *différentiel* are synonymous in this particular context.)

We repeat the process with the roles of T_1 and T_2 reversed. That is, we construct T_2' by replacing t_2 with t_1 in T_2 in order to verify whether the sense of t_1 in T_2' is synonymous with the sense of s in S_2. If the substitution test returns FALSE for either of the two target sentence pairs, we can conclude that the two multi-synsets that correspond to the senses of s in S_1 and S_2 must be different. Therefore, this instance of the WiC task is resolved as FALSE. However, if the substitution test returns TRUE for both pairs of sentences, we cannot immediately resolve the instance of the WiC task, because there could exist two (or more) multi-synsets that all contain s, t_1, and t_2. To complicate maters, this partial solution to the WiC task can only be applied if t_1 and t_2 have the same POS but different lemmas.

A complete theoretical solution can be obtained by considering translations in multiple languages. If the focus word s is not used in the same sense in S_1 and S_2, we would expect that in *some* language, the translations t_1 and t_2 will be different *and* not mutually replaceable in both sentences. This expectation is consistent with the speculation of Palmer et al. (2007) that translation into a sufficiently large set of language will eventually lexicalize every sense distinction. It is also supported by the findings of Bao et al. (2021) who found no evidence for the existence of universal colexifications, that is, pairs of concepts that are expressed by the same word in every natural language.

3.3 Multi-Synset Intersection

For each language F_i in the set of all natural languages $\mathcal{L}$, let t_1^i and t_2^i be the lexical translations of the focus word s in the first and second input sentences, respectively. Let T be the set consisting of the focus word, and all its lexical translations; that is $W = \{s\} \cup_{F_i} \{t_1^i, t_2^i\}$. Assuming access to a perfect universal multi-wordnet, we define the set C to be the set of multi-synsets that contain all words in T.

The size of C provides clues to the resolution of the WiC task. We need to consider three cases: $|C| = 0$, $|C| = 1$, and $|C| \geq 2$. With some caveats, these three cases roughly imply the following answers to the WiC task: FALSE, TRUE, and UNKNOWN, respectively. We discuss these three cases in turn.

If $|C| = 0$, then no single concept can be expressed by s and all its translations in T, according to the multi-wordnet. That is, there exist two translations of the focus word which cannot express the same concept, assuming the completeness of

the multi-wordnet. Therefore, the two focus tokens must correspond to distinct multi-synsets, implying FALSE.

If $|C| = 1$, there exists exactly one multi-synset that contains the focus word and all its translations. Therefore, it is possible, albeit not guaranteed, that the focus word in both source sentences is used in the sense that corresponds to that unique multi-synset. In order to be sure, we could apply the criss-cross method described in Section 3.2.

$|C| \geq 2$ would imply that there exist two concepts which are colexified (expressed by a single word) in all languages. Following Bao et al. (2021), we assume that universal collocations are at best extremely rare. Even if they exist at all, we could still apply the solution described in Section 3.2 to decide the WiC task. Of course, if we are considering translations into only a small number of languages, the possibility of $|C| \geq 2$ is much more likely. In fact, we observe $|C| = 3$ in our running example, because three different BabelNet multi-synsets contain the English focus word and its two French translations.

4 Methods

In this section we describe four methods based on the theoretical ideas in Section 3. All four methods rely on identifying lexical translations of the focus word in both source sentences. If the lexical translations cannot be recovered from the translated sentences for any of the target languages, all methods use the same backoff approach, which is to return FALSE for that test instance.

4.1 IDENT and CVAL

Our two simplest methods are IDENT and CVAL. IDENT is a baseline method which returns TRUE **iff** the lexical translations t_1 and t_2 have the same lemma and POS in all applicable target languages. CVAL is a method directly based on the cardinality of the set C as defined in Section 3.3. CVAL returns TRUE **iff** the translations of the focus word are identical in each language **and** $|C| > 0$.

4.2 Synonymy Check

We implement the substitution test as a heuristic *synonymy check* using dense contextualized embeddings. Such embeddings allow us to construct, for any word token in a given sentence, a vector in a continuous semantic space. The objective in designing such embeddings is that semantically similar tokens should have similar vectors, commonly measured by cosine similarity. Additional technical details of the embeddings are provided in Section 5.

Given a pair of sentences which differ only in the substitution of single word, we obtain dense contextualized embeddings of the distinguishing word in each sentence. We then calculate the cosine similarity between the two embeddings. If the similarity is greater than a threshold tuned on a development set, this is taken as an indication that replacing one of the distinguishing words with the other does not alter the meaning of the sentence, as the replacement word has the same meaning as the original word. This implementation of the substitution test is used as a subroutine by our remaining two methods.

4.3 SUB and CSUB

The SUB method attempts to apply the synonymy check to each pair of translated sentences T_1 and T_2 in each target language, without referring to the $|C|$ value. If the translations of the focus word in T_1 and T_2 differ, we create the sentences T_1' and T_2', as described in Section 3.2, and apply the synonymy check to (T_1, T_1') and (T_2, T_2'). SUB returns TRUE if the synonymy check succeeds for all target languages for which the translations t_1 and t_2 can be identified. The synonymy check trivially succeeds if t_1 and t_2 have the same POS and lemma; intuitively, tokens which translate the same way are likely to have similar meanings. If either application of the synonymy check fails, SUB returns FALSE. In summary, this method is similar to the IDENT method, except that the synonymy check is applied if the translations differ.

CSUB combines CVAL with SUB. The only difference with the SUB method is that the synonymy check is not applied when $|C| = 0$. This is because the lack of any common multi-synset in a complete perfect multi-wordnet is theoretically sufficient to exclude the possibility of the two source focus tokens having the same sense.

5 Experiments

In this section, we describe the application of our methods to the English development and test sets. We begin by specifying various implementation details. Next, we describe our development experiments, including results and error analysis. Finally, we present our results on the test set. While our

method is, in theory, applicable to any language, and even to cross-lingual subtasks, we focus exclusively on the English monolingual substask due to time and resource constraints.

5.1 Translation and Lemmatization

We use BabelNet (Navigli and Ponzetto, 2010, 2012) as our multi-wordnet; in particular, we make use of the BabelNet multi-synsets which are linked to Princeton Wordnet synsets. This allows us to exclude synsets that refer to named entities, rather than lexicalized concepts, to limit the impact of noise in BabelNet.

For translation, we use Google Translate, as it is fast and publicly available. In our analysis, we found the lexical translations obtained using Google Translate to be of generally high quality, which is important given our method's dependence on machine translation. We use French, Italian, and Russian as our languages of translation. The choice of the translation languages is based on the languages selected for the shared task, and also on the BabelNet coverage. French and Russian are two of the languages covered by the shared task. On the other hand, Italian seems to have the best BabelNet coverage among the non-English languages.

For lemmatization, we use TreeTagger (Schmid, 1999, 2013), with pre-trained lemmatization models for the source and all target languages. We lemmatize the bitexts to improve the quality of the word alignment.

5.2 Word Alignment

Following lemmatization, we align each input sentence with its translation in each target language. To improve the quality of our unsupervised alignment, we obtain a large sentence-aligned parallel corpus (*bitext*) in the source and target languages. We then append to the bitext all of the lemmatized input sentences, and all of their lemmatized language translations. Finally, we apply an unsupervised knowledge-based alignment algorithm to the augmented bitext, and, for each sentence, identify the word or phrase in the translated sentence corresponding to the source focus word. Once each input sentence is aligned with its translation, we extract the lemmas aligned with each focus word token. These are the lexical translations of the focus word for this language.

To carry out the alignment, we use BabAlign (Luan et al., 2020), a state-of-the-art knowledge-based aligner. BabAlign leverages translation information from BabelNet to create synthetic training data and post-process the alignment produced using a base unsupervised alignment method. Specifically, we use FastAlign (Dyer et al., 2013) as the base aligner. When aligning input sentences with translations, we concatenate the sentences and their translations with the OpenSubtitles bitext (Lison and Tiedemann, 2016) for the corresponding language pair. For each language pair, we use the first 1M sentences of the OpenSubtitles bitext.

5.3 Contextual Embeddings

To obtain contextual representations for the purposes of deciding the substitution check, we use BERT (Devlin et al., 2019), a deep neural architecture trained with the masked language model. We chose BERT because it has been proven to capture the semantics of a word in context (Coenen et al., 2019). The context is the sentence containing the focus word. Specifically, we use cased multilingual BERT to generate contextualized embedding of focus words by summing up the last four hidden layers of the BERT model. This choice was based on the results achieved by Devlin et al. (2019) in the named entity recognition task, and by Soler et al. (2019) in the SemDeep-5 WiC shared task.[1]

We use cased multilingual BERT embeddings with 768 dimensions, 12 layers, 12 attention heads, and 179M parameters. To implement the substitution check, we generate contextualized embeddings of the translations of the focus tokens, and their substitutes, by summing the last four hidden layers of the BERT model. Since BERT uses sub-tokens to generate embeddings, we analyzed the impact of two different sub-token selection techniques for predicting word similarity: using only the first sub-token, and using the mean over all the sub-tokens. In our development experiments, we found that the former yielded better results. Therefore, only the first sub-token is used to create contextualized embeddings for the substitution method.

5.4 Development Results

Table 1 shows the results of our development experiments. The baseline translation identity method IDENT does surprisingly well, outperforming both methods based on intersecting sets of multi-synsets, CVAL and CSUB. Indeed, these methods tend to suffer accuracy degradation as more languages of translation are added. We speculate that this is due

[1] https://www.dfki.de/ declerck/semdeep-5/challenge.html

Lang.	FR	IT	RU	ALL
IDENT	**59.6**	**58.1**	**57.1**	59.7
CVAL	58.9	57.6	54.3	55.5
SUB	59.3	58.0	55.6	**60.8**
CSUB	59.2	57.8	54.3	54.1

Table 1: MCL-WiC accuracy (%) on the En-En dev set with different methods and languages of translation.

Lang.	FR	IT	RU	ALL
IDENT	55.8	**58.9**	**61.0**	61.1
CVAL	54.8	55.6	56.0	55.2
SUB	**56.1**	57.6	60.6	**63.2**
CSUB	55.2	55.2	55.8	55.7

Table 2: MCL-WiC accuracy (%) on the En-En test set with different methods and languages of translation.

to these methods being more vulnerable to noise (errors or omissions) in the multi-wordnet and in the extraction of lexical translations. However, the best performing method is SUB, which also shows improvement when combining all three languages of translation. Thus, it also shows the most promise for further improvement by adding additional languages.

Our error analysis suggests that there are three principal causes of errors. First, translation may be non-literal. For example, in one instance, the adverb "unevenly" is translated into French as the adjective "inégale" ("unequal"), leading to a false negative. Second, distinct but synonymous translations may lead to false positives. In one instance, the focus word "stain" is translated as "souillé" in one sentence and "tachée" in the other. The focus tokens have distinct meanings, reflected in their distinct translations, "stain on a reputation" versus "stain on a surface". However, the translations pass the BERT-based synonymy check, since they can be synonymous in some contexts. Finally, in some cases, distinct senses of a word may nevertheless translate the same way. For example, in one instance, the focus word "superior" was used in two distinct meanings. Both these meanings can be expressed by the French word "supérieur", and indeed, "superior" was translated as "supérieur" in both sentences, resulting in a false positive.

5.5 Test Results and Discussion

Table 2 shows our results on the test data. Consistent with our development experiments, the SUB method achieves the best performance with the combination of all three languages. The IDENT method once again performs surprisingly well despite its simplicity, outperforming the more complex CVAL and CSUB methods. Different from the development experiments, when only one language of translation is used, Russian yields substantially better performance compared to French or Italian across all four methods, and Italian likewise yields better performance than French.

Table 3 gives additional details for the results of the SUB method. For each of the three languages, and the combination of all three, we provide the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as well as the accuracy. We observe that using multiple languages of translation results in a substantial reduction in false positives, at the possible expense of an increase in false negatives, while maintaining an overall higher accuracy.

Lang.	TP	TN	FP	FN	Accuracy
FR	369	192	308	131	56.1
IT	376	200	300	124	57.6
RU	327	279	221	173	60.6
ALL	339	293	207	161	63.2

Table 3: Detailed breakdown of the results of our best performing method, SUB.

6 Conclusion

Overall, our results provide a solid proof-of-concept for the utility of multilingual translation for the WiC task. While not competitive with state-of-the-art supervised methods, our results empirically verify the hypothesis that translations convey semantic information, and that this phenomenon has applications in lexical semantics. The IDENT and SUB methods consistently benefit from translation into multiple languages, and this result generalizes to unseen test data.

Acknowledgements

We thank the organizers of the shared task for their effort. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Luis Espinosa Anke, Thierry Declerck, Dagmar Gromann, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2019. Proceedings of the 5th workshop on semantic deep learning (SemDeep-5). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*.

Carlos Santos Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Ivan Vulić, and Mohammad Taher Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 36–49.

Hongchang Bao, Bradley Hauer, and Grzegorz Kondrak. 2021. On universal colexifications. In *Proceedings of the 11th Global Wordnet Conference (GWC2021)*, pages 1–7.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Vigas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *arXiv preprint arXiv:1906.02715*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Christiane Fellbaum. 1998. WordNet: An on-line lexical database and some of its applications. *MIT Press*.

Goran Glavaš, Ivan Vulić, Anna Korhonen, and Simone Ponzetto. 2020. SemEval-2020 task 2: Predicting multilingual and cross-lingual (graded) lexical entailment. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Bradley Hauer, Amir Ahmad Habibi, Yixing Luan, Arnob Mallik, and Grzegorz Kondrak. 2020. UAlberta at SemEval-2020 task 2: Using translations to predict cross-lingual entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 263–269, Barcelona (online). International Committee for Computational Linguistics.

Bradley Hauer and Grzegorz Kondrak. 2020a. One homonym per translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7895–7902.

Bradley Hauer and Grzegorz Kondrak. 2020b. Synonymy = translational equivalence. *arXiv preprint arXiv:2004.13886*.

Bradley Hauer, Yixing Luan, and Grzegorz Kondrak. 2019. You shall know the most frequent sense by the company it keeps. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 208–215. IEEE.

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. European Language Resources Association.

Qianchu Liu, Diana McCarthy, and Anna Korhonen. 2020. Towards better context-aware lexical semantics: Adjusting contextualized representations through static anchors. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4066–4075.

Daniel Loureiro and Alípio Jorge. 2019a. 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.

Daniel Loureiro and Alipio Jorge. 2019b. LIAAD at SemDeep-5 challenge: Word-in-Context (WiC). *arXiv preprint arXiv:1906.10002*.

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving word sense disambiguation with translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval-2021)*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. 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.

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.

Lucas RC Pessutto, Tiago de Melo, Viviane P Moreira, and Altigran da Silva. 2020. BabelEnconding at SemEval-2020 task 3: Contextual similarity as a combination of multilingualism and language models. *arXiv preprint arXiv:2008.08439*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3528–3539.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Helmut Schmid. 2013. Probabilistic part-of-speech tagging using decision trees. In *New methods in language processing*, page 154.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. LIMSI-MULTISEM at the IJCAI SemDeep-5 WiC challenge: Context representations for word usage similarity estimation. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 6–11.