SzegedAI at SemEval-2021 Task 2: Zero-shot Approach for Multilingual and Cross-lingual Word-in-Context Disambiguation

Gábor Berend
Institute of Informatics
University of Szeged
Árpád tér 2., Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract

In this paper, we introduce our system that we participated with at the multilingual and cross-lingual word-in-context disambiguation SemEval 2021 shared task. In our experiments, we investigated the possibility of using an all-words fine-grained word sense disambiguation system trained purely on sense-annotated data in English and draw predictions on the semantic equivalence of words in context based on the similarity of the ranked lists of the (English) WordNet synsets returned for the target words decisions had to be made for. We overcame the multi,-and cross-lingual aspects of the shared task by applying a multilingual transformer for encoding the texts written in either Arabic, English, French, Russian and Chinese. While our results lag behind top scoring submissions, it has the benefit that it not only provides a binary prediction whether two words in their context have the same meaning, but also provides a more tangible output in the form of a ranked list of (English) WordNet synsets irrespective of the language of the input texts. As our framework is designed to be as generic as possible, it can be applied as a baseline for basically any language (supported by the multilingual transformed architecture employed) even in the absence of any additional form of language specific training data.

1 Introduction

A major obstacle in solving word sense disambiguation (WSD) problems in a supervised manner is the scarcity of annotated training corpora. As the construction of high quality sense-annotated training data can be extremely labor-intensive and difficult (Gale et al., 1992), the Word-in-Context (WiC) disambiguation task was recently proposed by Pilehvar and Camacho-Collados (2019) as a surrogate for the traditional WSD problem. While in the traditional fine-grained WSD setting, the aim is to assign a precise and often nuanced meaning to a word in its context according to some sense inventory, WiC is framed as a binary classification problem, where the task is to decide whether two target words originating from a pair of input sentences have the same meaning. This kind of binary decision can also be made in the absence of a nuanced sense inventory, making the annotation process less demanding and also more suitable across languages (Raganato et al., 2020).

In this paper, we analyze the utilization of multilingual transformer-based language models for performing both multi-lingual and cross-lingual WiC in the zero-shot setting, by employing nothing but English sense annotated training data and utilizing the model predictions in a transductive model that is capable of performing zero-shot WSD and WiC disambiguation for any language that is supported by the multilingual transformer encoder model that gets employed.

Loureiro and Jorge (2019) showed that a simple, nearest neighbor approach relying on contextual word embeddings can achieve impressive WSD results in English. In our follow-up work (Berend, 2020), we demonstrated, how sparse contextualized word representations can be exploited for obtaining significant improvements over the LMMS approach introduced by Loureiro and Jorge (2019). Our shared task participation was focused on comparing the two techniques in a zero-shot multilingual and cross-lingual WiC evaluation setting.

2 System overview

At the core of our multi,-and cross-lingual WiC systems, we employed fine-grained WSD systems, originally intended to solely handle English texts. The two models that we employed were the LMMS (Loureiro and Jorge, 2019) and the S-LMMS (Berend, 2020) approaches. We dub the

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 169–174
Bangkok, Thailand (online), August 5–6, 2021.

latter solution as S-LMMS, highlighting its resemblance to the LMMS approach and the fact that it operates with *sparse* contextualized word representations. Both LMMS and S-LMMS requires sense-labeled training data for constructing their respective fine-grained WSD models.

We provide a brief overview of the two approaches and encourage readers interested in more details to read the original papers (Loureiro and Jorge, 2019; Berend, 2020) introducing them. LMMS and S-LMMS both has in common, that they encode the inputs with a transformer model (BERT-large). LMMS constructs a prototype vector for each English synset based on the BERT-encoded vectors of the sense-annotated training data and the actual contents of the English WordNet glosses. For a given token in its context, LMMS takes its BERT-encoded contextualized vector and finds the nearest synset prototype for determining its sense.

The way S-LMMS differs from LMMS is that it additionally incorporates a sparsity inducing dictionary learning step, which turns the contextualized word representations into a sparse format, i.e., to such vectors that contain a high fraction ($> 90\%$) of zero coefficients. Additionally, the methodology for creating the synset prototype vectors has substantial differences between the two approaches, as LMMS uses the actual contextualized embeddings pertaining to a certain synset as prototypes, whereas S-LMMS distills a vectorial representation to each synset based on an information theoretic measure.

The important technical change that we performed over the previously described fine-grained WSD models, so that they can be employed in the cross-lingual setting, is that we replaced the BERT-large encoders that the LMMS and S-LMMS models use by default to the XLM-RoBERTa-large (Conneau et al., 2020) architecture. We shall refer to the variants of LMMS and S-LMMS that were obtained by relying on XLM-RoBERTa as an encoder as opposed to BERT-large as mLMMS and mS-LMMS, owing to the multilingual nature of XLM-RoBERTa. We used the transformers library (Wolf et al., 2020) for obtaining the contextualized multilingual embeddings for our experiments.

When performing fine-grained WSD in English, one can simply restrict the scope of predicting the most likely synset for some word to those that are deemed viable for a given word in WordNet. Additionally, one can also filter the synsets over which the prediction is performed, based on the part-of-speech category of a word in question. With these heuristics, it is possible to reduce the number of synsets that a word can belong to a few dozens of synsets even for the most ambiguous cases.

In order to test a solution that is as generic as possible, we did not integrate any of these heuristics into our framework, meaning that our models returned a ranked list over *all* the 117,659 English WordNet synsets to any word from some sentence. This way, our solution can also work basically any language (supported by the multilingual transformer employed), even in the absence of a multilingual sense-inventory resource such as BabelNet (Navigli and Ponzetto, 2010) and also when we have no access to the part-of-speech information, nor to a part-of-speech tagger for some language. These design choices ensures that we are able to handle a much wider range of languages as if we decided otherwise. To this end, we regard our approach a particularly good fit being used as a baseline for WSD related evaluations involving low-resource languages.

As mentioned previously, our `⋆LMMS` models assigned a ranked list of 117,659 English synsets to every target word irrespective of the language of the sentence it was written in. Since the ranking of the synsets for a given word was performed over all the synsets of WordNet, it would be too restrictive to expect that words with identical meaning should be assigned the exact same most likely English synset. To this end, we measured the similarity for a pair of ranked lists that a model returned for a pair of words in their contexts and decided about the semantic equivalence of the two words based on that similarity score. As the similarity scores calculated for the ranked lists of synsets that fit those pairs of words that have the same meaning are expected to be higher on average, we decided to determine a threshold for the similarity scores of the ranked lists above which we predicted the two words to have the same meaning, and to have a different meaning otherwise.

We experimented with three strategies for measuring the similarity of two ranked synset lists for a pair of words. Let S_1 and S_2 refer to the ranked lists of WordNet synsets assigned to two words. As the bottom of the ranking is arguably not as meaningful as its top-ranked elements, we decided to formulate $S_1^{(100)}$ and $S_2^{(100)}$. These ranked lists

differed from S_1 and S_2 in that they contained their top 100-ranked elements, respectively.[1]

Since we only focus on the highest ranked synsets from S_1 and S_2, it is almost sure that certain element from $S_1^{(100)}$ are not included in $S_2^{(100)}$, and vice versa. As such, the usage of standard rank correlation scores would be inconvenient for measuring the similarity between ranked lists $S_1^{(100)}$ and $S_2^{(100)}$. One motivation behind the introduction of ranking-biased overlap (RBO) (Webber et al., 2010) was particularly this, i.e. to provide such a distance metric that is capable of operating between non-conjoint rankings. RBO is an overlap-based metric, that can operate over such rankings when the ranked elements themselves are not totally identical. To this end one of our metric for measuring the similarity between $S_1^{(100)}$ and $S_2^{(100)}$ was based on the RBO metric.

Our other approach for measuring the similarity of ranked lists $S_1^{(100)}$ and $S_2^{(100)}$ was to simply take their Jaccard similarity, i.e. the fraction of the size of their intersection and the elements in their union. As a third approach, we calculated the harmonic mean of the mean reciprocal rank (MRR) of the highest ranked synset from $S_1^{(100)}$ in the ranked list $S_2^{(100)}$ and similarly, that of the highest ranked synset from $S_2^{(100)}$ in $S_1^{(100)}$. We then based our predictions with the similarity scores calculated by either of the above manner.

Instead of using some supervised approach, we determined a threshold for the similarity score for a pair of ranked synset lists $S_1^{(100)}$ and $S_2^{(100)}$, above which we predicted that the words they got assigned to had identical meaning. We determined this threshold in a transductive manner, without using any of the labeled training or development set sentence pairs at all. For the cross-lingual evaluation it would have been impossible at the first place, as no annotated pairs of sentences were released during the shared task.

We used expectation maximization for determining the similarity threshold above which we predicted a pair of words to have the same meaning. That is, we took all the similarity scores that we calculated for a certain test set based on the $S_1^{(100)}$ and $S_2^{(100)}$ ranked synset lists, and fitted a Gaussian Mixture Model over the similarity scores. That way, we managed to fit a Gaussian distribution for the similarity scores of pairs of words with identical and different meanings. We identified the fitted Gaussian distribution with the higher expected value to be the one that corresponds to the distribution of similarity scores for those words that have identical meaning. As expectation maximization algorithms are prone to find local optima, we initialized each model 100 times and chose the one which resulted in the best log-likelihood score. Our decisions for a particular test sample was then based on the density functions on the similarity scores of the two classes determined by the best fitting model.

3 Experiments

We tested our approach on both the multilingual and the cross-lingual subtasks of the shared task (Martelli et al., 2021). The multilingual test sets consisted of sentence pairs that were written in the same language (either Arabic, English, French, Russian or Chinese), whereas, an input was comprised of an English and a non-English (either Arabic, French, Russian or Chinese) sentence for the cross-lingual scenario.

The fine-grained WSD model that we built our system on was trained over English sense-annotated training data. We used two sources of training signal, the SemCor dataset as well as the Princeton WordNet Gloss Corpus (WNGC), which has been shown to improve fine-grained WSD results (Vial et al., 2019; Berend, 2020). Unless stated otherwise, we used these three sources of sense-annotated training data for obtaining our $\star$LMMS models.[2]

3.1 Monolingual all-words WSD experiments

We first evaluated LMMS and S-LMMS models on standard fine-grained all-words disambiguation data included in the unified evaluation framework from (Raganato et al., 2017). What we were interested here is the change in the standard WSD performance of these systems when replacing the English specific BERT-large model that LMMS and S-LMMS originally employ to XLM-RoBERTa-large. At this point we evaluated our fine-grained WSD performance in terms of F-score over the concatenation of the five standard evaluation benchmarks from SensEval2 (Edmonds and Cotton, 2001), SensEval3 (Mihalcea et al., 2004), SemEval 2007 Task 17 (Pradhan et al., 2007), SemEval 2013 Task 12

[1] Experiments with different thresholds (10, 25, 50, 250 and 500) also provided similar results that we omit for brevity.

[2] Our source code can be found at https://github.com/begab/sparsity_makes_sense

Layer(s) used	LMMS	S-LMMS
21	0.758	**0.790**
22	0.763	**0.785**
23	0.760	**0.786**
24	0.745	**0.780**
21-24	0.757	**0.788**

(a) BERT-large

Layer(s) used	mLMMS	mS-LMMS
21	0.702	**0.757**
22	0.692	**0.753**
23	0.679	**0.749**
24	0.648	**0.728**
21-24	0.692	**0.754**

(b) Using XLM-RoBERTa-large

Table 1: Comparison of the model performances towards fine-grained WSD using the standard benchmark from (Raganato et al., 2017) (consisting of the concatenated test sets of the SensEval2-3 and the SemEval 2007, 2013 and 2015 shared tasks on fine-grained WSD), when using different layers from different transformer models and model variants ⋆LMMS.

(Navigli et al., 2013), SemEval 2015 Task 13 (Moro and Navigli, 2015). This test set consisted of 7,253 English test cases in total.

Table 1 includes our results using the four different models that were using different layers from the transformer model that was employed for encoding the input texts. As expected, replacing the English specific transformer model to a multilingual encoder resulted in a decreased performance, however, the overall decrease was not very severe. Comparison of the results in Table 1a and Table 1b reveals that the performance of S-LMMS is less affected by the integration of the multilingual RoBERTa model in place of the English-only BERT model for encoding. Additionally, using the encodings from the 21th layer of the transformer models seem to provide a slight edge over the utilization of the concatenation of the last four layers irrespective of the encoder and the specific WSD model used. To this end, we participated in the shared task-related with such ⋆LMMS models that were using the contextualized word representations from the 21th layer alone, as opposed to the average of the last four layers.

3.2 Evaluation on the shared task data

In Table 2, we list those test scores that we obtained by differently configured versions of our architecture. Our results span the different strategies for performing all-words fine-grained WSD (mLMMS/mS-LMMS) and different strategies for calculating the similarity between two ranked list of most likely synsets assigned to the test words (Jaccard/MRR/RBO) as described earlier in Section 2.

We can see from Table 2 the same phenomenon as for our monolingual fine-grained WSD evaluations in Table 1, i.e., the mS-LMMS approach had a clear advantage over LMMS for both the multilingual and the cross-lingual evaluation settings.

Regarding the effects of choosing different ways to calculate the similarity scores between a pair of ranked lists of synsets, the application of the Jaccard similarity and the RBO metric-based similarity seems to perform very similarly, with the mean reciprocal rank based similarity scoring slightly underperforming the other two alternatives. Overall, the results seem to be balanced over the languages, with the choice of the fine-grained WSD system being more influential to the final results as the choice of the similarity calculation between the ranked lists of synsets returned by them to a pair of test words.

For training our ⋆LMMS models, we decided to experiment with the integration of a recent source of sense tagged training dataset, UWA (Loureiro and Camacho-Collados, 2020), which is a sense-annotated corpus containing unambiguous words from Wikipedia and OpenWebTex. We relied on the recommended version of the UWA corpus which contains 10 example sentences for each unambiguous word. By expanding the number of sense annotated training text, it becomes possible to increase the coverage of the fine-grained WSD systems. We investigated the downstream effects for our WiC system of extending the amount of sense annotated training data used by our fine-grained WSD systems.

Our evaluation results over the same set of models as in Table 2, with the only difference that we additionally used the UWA10 sense-annotated corpus for creating our all-words WSD models are included in Table 3. This additional training corpus was not always helpful, however, increased our

	Jaccard		MRR		RBO	
	mLMMS	mS-LMSS	mLMMS	mS-LMMS	mLMMS	mS-LMSS
ar	60.0	61.4	**62.1**	60.7	59.2	59.5
en	62.6	67.2	**70.6**	70.4	62.6	66.1
fr	62.1	66.6	62.4	60.9	60.7	**66.9**
ru	58.9	67.1	63.9	66.6	56.6	**67.3**
zh	55.9	63.8	56.0	63.8	56.7	**64.6**
avg.	59.9	**65.2**	63.0	64.5	59.2	64.9

(a) Multilingual results

	Jaccard		MRR		RBO	
en-*	mLMMS	mS-LMSS	mLMMS	mS-LMMS	mLMMS	mS-LMSS
ar	59.9	**66.3**	59.1	64.4	61.3	62.2
fr	61.2	63.9	59.5	63.1	59.6	**64.6**
ru	63.7	**66.4**	61.2	60.2	62.7	65.9
zh	64.2	65.3	51.5	65.6	62.9	**66.3**
avg.	62.3	**65.5**	57.8	63.3	61.6	64.8

(b) Cross-lingual results

Table 2: The effects of applying different similarity measures (Jaccard/MRR/RBO) to the different fine-grained WSD approaches (mLLS/mS-LMMS) integrated into our zero-shot multilingual and cross-lingual WiC framework.

	Jaccard		MRR		RBO	
	mLMMS	mS-LMSS	mLMMS	mS-LMMS	mLMMS	mS-LMSS
ar	58.2	64.1	65.0	**66.9**	58.0	63.8
en	62.4	68.2	69.3	**70.2**	63.0	68.7
fr	62.4	67.8	60.1	61.6	62.8	**68.3**
ru	57.9	**68.7**	64.8	67.7	60.5	66.2
zh	51.3	63.0	**64.9**	63.9	53.3	64.2
avg.	58.4	**66.4**	64.8	66.1	59.5	66.2

(a) Multilingual results

	Jaccard		MRR		RBO	
en-*	mLMMS	mS-LMSS	mLMMS	mS-LMMS	mLMMS	mS-LMSS
ar	59.5	**65.8**	60.2	64.7	58.8	63.2
fr	60.6	**64.9**	62.0	59.8	60.5	64.5
ru	63.0	65.9	61.8	59.6	62.7	**68.3**
zh	60.3	66.0	53.1	65.5	60.9	**67.1**
avg.	60.9	65.7	59.3	62.4	60.7	**65.8**

(b) Cross-lingual results

Table 3: The effects of incorporating the UWA10 sense-annotated corpus during the training phrase of our fine-grained English WSD model that served as a basis of our WiC architecture.

average accuracy by a slight ($\approx 1\%$) margin.

4 Conclusions

In this paper, we introduced our cross,-and multilingual WiC framework that we approached from an all-words fine-grained word sense disambiguation perspective. As such, our model not only provides a yes or no answer for a pair of words in their contexts, but also provides a more tangible explanation for it in the form of the similarity between the ranked lists of English WordNet synsets assigned to the target words.

During the design of our approach, we made such choices that would make our framework conveniently applicable to new languages without the need for any training data. Although the results of our framework lags behind the top performing systems, due to of its convenient applicability to new languages and the fact that practically no additional training data is required for applying it to new and possibly low-resourced languages, we think it can provide an easy to use baseline in further WiC-related research efforts.

Acknowledgments

The research presented in this paper was supported by the Ministry of Innovation and the National Research, Development and Innovation Office within the framework of the Artificial Intelligence National Laboratory Programme. The author is grateful for the fruitful discussions with Tamás Szakálos whose research was supported by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002. The project has been supported by the European Union and co-funded by the European Social Fund.

References

Gábor Berend. 2020. Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8498–8508, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David

Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.

Daniel Loureiro and Jose Camacho-Collados. 2020. Don't neglect the obvious: On the role of unambiguous words in word sense disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3514–3520, Online. Association for Computational Linguistics.

Daniel Loureiro and Alípio Jorge. 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.

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)*.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The SENSEVAL-3 english lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual 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 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameer S. Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, srl and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 87–92, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. 2019. Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In *Global Wordnet Conference*, Wroclaw, Poland.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.