FII_CROSS at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation

Ciprian Bodnar[1], Andrada Tapuc[1], Cosmin Pintilie[1], Daniela Gîfu[1,2], Diana Trandabăț[1,3]
[1]Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iasi
[2]Institute of Computer Science, Romanian Academy - Iasi Branch
[3]Imagination Play SRL
{ioan.bodnar, andrada.tapuc, cosmin.pintilie, daniela.gifu, dtrandabat}@info.uaic.ro

Abstract

This paper presents a word-in-context disambiguation system. The task focuses on capturing the polysemous nature of words in a multilingual and cross-lingual setting, without considering a strict inventory of word meanings. The system applies Natural Language Processing algorithms on datasets from SemEval 2021 Task 2, being able to identify the meaning of words for the languages Arabic, Chinese, English, French and Russian, without making use of any additional mono- or multilingual resources.

1 Introduction

The computational task of disambiguating a word in the context of its sentence is still a very challenging topic facing natural language processing (NLP). In this study, we refer to word meaning that requires a multidisciplinary approach for its detection. From sense-based and contextualized embeddings, all tries are aimed at providing an understanding of words in context. We notice that evaluating such approaches is not easy. For instance, traditional Word Sense Disambiguation (WSD) fails to test latent representations unless these are linked to explicit sense inventories such as WordNet (Matusevych, 2016) or BabelNet (Navigli and Pozetto, 2012; Luan *et al.*, 2020). To resolve the problem of disambiguation for both lingual dimensions, we tried to use a combination of well-known algorithms to provide an optimal system.

The legitimate research questions this paper intend to answer: *Is Word-in-Context Disambiguation a barrier for NLP techniques?*

The approach we propose in this paper investigates two models of cross-lingual word embeddings, comparing them to the shared-translation effect and the cross-lingual coactivation effects of false and true friends (cognates) found in human language. We find that the similarity structure of the cross-lingual word embeddings space yields the same effects as the human bilingual lexica (Merlo and Rodriguex, 2019). Research on bilingual lexica has uncovered fascinating interactions between the L1 (native language) and L2 (second language) lexica showing that both production and comprehension coactivate lexical items in both languages, indicating that bilinguals store lexical representations from their native and their second language in the same space.

The rest of the paper is organized as follows: section 2 describes the literature related to sense disambiguation, section 3 presents the dataset and method of this study, section 4 resumes the results of the conducted experiments, followed by section 5 with the conclusions and discussions about how to increase the accuracy.

2 Background

This topic has attracted significant attention in recent years, evidenced by increasing number of workshops (e.g., SemEval-2013 Task 10: Cross-lingual world Sense Disambiguation - CLWSD). Participating in such competitions is especially attractive since teams have thus access to labeled data.

For binary tasks, as the case of this competition, there are many computational models to be used in detecting the right word sense.

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 787–792
Bangkok, Thailand (online), August 5–6, 2021.

The recent advancements in corpus linguistics technologies, as well as the availability of more and more textual data, encourage many researchers to take advantage of comparable and parallel corpora to address different NLP tasks. Work on this topic is however highly subjective and biased. In general, the methods are based on Bag of Words features, usually normalized with *tf*idf* or character n-grams features for stylistic purposes.

Most approaches are supervised methods which can be classified into different methods:

(1) *regression*, based on the embeddings in one language using a leastsquares objective (Dinu et al., 2015; Artexe et al., 2018);

(2) *orthogonal*, based on the embeddings in one or both languages under the constraint of the transformation (Zhang et al., 2016; Smith et al., 2017);

(3) *canonical*, based on the embeddings in both languages to a shared space, using canonical link extension of it (Lu et al., 2015).

Also, several systems use an approach similar to ours' in considering Sent2vec the main algorithm (Pagliardini et al., 2018). Other systems used a *maxent* classifier trained over local context or even a KNN (*K-Nearest Neighbors*) classifier to solve the CLWSD (*Cross-Lingual Word Sense Disambiguation*) task. One of the interesting approaches was using machine translation (Baker et al., 1993). Although the winning systems for the CLWSD task used different approaches (statistical machine translation and classification algorithms), they also only used a parallel corpus to extract disambiguating information, while not using external resources such as WordNet. As a consequence, their system is very flexible and language-independent.

The topic of multilingual and cross-lingual disambiguation has attracted significant attention in recent years (Akyürek et al., 2020), with approaches ranging from learning effective vector representations (Loureiro and Jorge 2019, Scarlini et al., 2020) to infusing neural networks with knowledge graph information (Bevilacqua and Navigli 2020).

Our approach is more focused on recent research on the bilingual lexicon, which uncovered fascinating interactions between the lexica of the native language and that of the second language in bilingual speakers. Thus, it has been found that the lexicon of the underlying native language affects the organization of the second language (Riley et al., 2020). In that spirit, our system includes distributed representations to disambiguate words in context.

3 Datasets and Methods

The dataset for the SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation is detailed in the task description paper (Martelli et al., 2021).

Language	Total words	Training words	Testing words
Arabic	20000	15000	5000
Chinese	16000	12000	4000
English	24500	16500	8000
French	22000	15500	6500
Russian	20000	14500	5500

Table 1: Corpus statistics

3.1 Dataset

The dataset (Table 1), released in JSON format and divided into .data and .gold files, had the following composition: training, development and test subsets for multilingual and cross-lingual settings.

The data files contain the following information:

(1) unique id of the pair;

(2) target lemma;

(3) part of speech;

(4) first sentence;

(5) second sentence;

(6) start and end indices of the target word occurring in the first and second sentence.

The training data set contains 8000 entries for each multilingual language and 8000 entries for cross-lingual language combinations, and in the test dataset there are 1000 entries for each. The .gold files contain the following information, as exemplified below:

- unique id of the pair
- tag (binary, either T/F)

```
{
   "id": "training.en-en.0",
   "tag": "F"
},
```

We used NLTK to tokenize the sentences and removed the stop words, only keeping the lemmas in sentence1 and sentence2 respectively, and the

gold tag. After the transformation, the data in the training files look as exemplified below:

```
  {
    "sentence1":              "context
coordination integration Bolivia hold
key   play   process   infrastructure
development ",
    "sentence2": "school water needed
girl sent fetch taking time away study
play ",
    "lemma": "play",
    "tag": "F"
  },.
```

3.2 Methods

We considered that our system will first solve the multilingual part, followed by the cross-lingual part. Therefore, we propose the specific architecture presented in Figure 1.

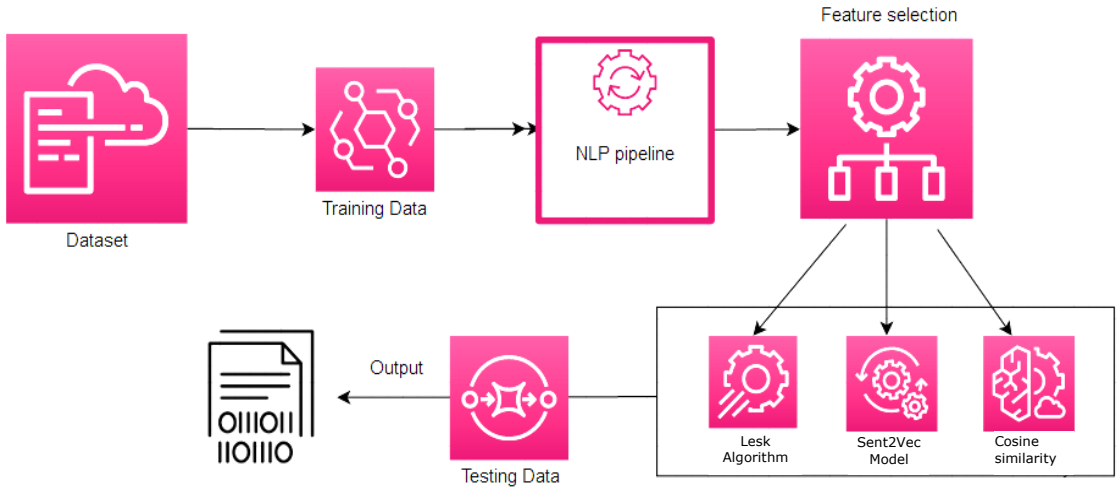

Figure 1. The System Architecture

Sent2Vec presents a simple but efficient unsupervised method to train distributed representations of sentences. It can be thought of as an extension of FastText and Word2vec (CBOW) to sentences. The sentence embedding is defined as the average of the source word embeddings of its constituent words. This model is furthermore augmented by learning source embeddings for both unigrams and various n-grams of words occurring in sentences and averaging the n-gram embeddings along with the words (Pagliardini et al., 2018). Thus, in our output, we have the initial tag from the labelled data, along with the score obtained through Sent2Vec.

As baseline, we used the Lesk algorithm. Thus, we extracted the definition of the queried target word from the first and the second sentence, respectively. Finally, we compared the definitions of the target word in the two contexts, thus reaching the *True* or *False* tag. Lesk output before comparing definitions is presented below. Using NLTK-WORDNET, we extracted all synonyms for a target word.

```
Sentence 1: In that context of
coordination and integration, Bolivia
holds a key play in any process of
infrastructure development.

-Sense:Synset('play.v.02')
-Definition:act or have an effect in
a specified way or with a specific
effect or outcome
```

```
Sentence 2: A musical play on the
same subject was also staged in
Kathmandu for three days.

-Sense:Synset('play.v.28')
-Definition:discharge or direct or
be discharged or directed as if in a
continuous stream
```

Our final approach was to combine the Lesk algorithm, along with Sent2Vec and vector cosine (Bojanowski et al., 2017). The cosine similarity is computed for each pair of sentences in our input. The pipeline used cross-lingually aligned versions of *fasttext* word vectors.

After running all modules, we obtained two scores, given by the cosine similarity and Sent2Vec

respectively, and a tag (*T* or *F*) provided by the Lesk algorithm. In order to apply an integrative approach, we transformed the tags from the Lesk algosithm into a 3rd score: if the tag is T the Lesk score became 0.9, while if the tag is F, the score became 0.3. These values were determined empirically, after several tests with different weights. Below are the scores obtained for the pair of sentences discussed above.

```
sent2vec score: 0.0912621021270752
lesk score: 0.3
cosine score: 0.14142135623730948
gold tag: F
```

In order to establish a ranking on the three scores and their influence on the final tag, we tested and analyzed several combinations of weights, as follows:

```
Score1 – 30%sent2vec*10 + 30%lesk +
40%cosine*10: 0.9294717313304636
Score2 – 30%sent2vec*10 + 20%lesk +
50%cosine*10: 1.0408930875677729
Score3 – 30%sent2vec*10 + 40%lesk +
30%cosine*10: 0.818050375093154
 Score4 – 40%sent2vec*10 + 20%lesk
+ 40%cosine*10: 0.9907338334575388
```

After a detailed error analysis in which we compared the gold tags from the development corpus with those issued in each of our combinations, we decided that the final tags be assigned according to the score brought by the formula (1):

```
(1) FiiCros_formula = 20%*Lesk + 30%
* Sent2Vec*10 + 50% Cosine *10
```

4 Results

The results for each individual task (Precision, Recall and F1-score) using the specific test dataset are presented: for multilingual test dataset (Table 2) and for crosslingual dataset (Table 3). The baseline identified 4483 correct tags out of 8000 inputs. After fine tuning the weights of the combination of our algorithms with the final formula discussed above our system reached 5760 correct tags out of 8000 inputs.

Model	Precision	Recall	F1-score
Lesk Baseline EN-EN	56%	65%	60,17%
Lesk + Sent2Vec + Cosine Vectorial EN-EN	72%	68%	69,94%
Lesk Baseline FR-FR	52%	61%	56,14%
Lesk + Sent2Vec + Cosine Vectorial FR-FR	70%	66%	67,94%
Lesk Baseline AR-AR	50%	59%	54,13%
Lesk + Sent2Vec + Cosine Vectorial AR-AR	68%	64%	65,94%
Lesk Baseline ZH-ZH	51%	60%	55,14%
Lesk + Sent2Vec + Cosine Vectorial ZH-ZH	69%	65%	66,94%
Lesk Baseline RU-RU	53%	62%	57,15%
Lesk + Sent2Vec + Cosine Vectorial RU-RU	71%	67%	68,94%

Table 2. Experimental Results for multilingual test dataset

Model	Precision	Recall	F1-score
Lesk Baseline EN-AR	53%	61%	56,71%
Lesk + Sent2Vec + Cosine Vectorial EN-AR	70%	66%	67,94%
Lesk Baseline EN-FR	49%	58%	53,12%
Lesk + Sent2Vec + Cosine Vectorial EN-FR	67%	64%	65,46%
Lesk Baseline EN-ZH	44%	55%	48,88%
Lesk + Sent2Vec + Cosine Vectorial EN-ZH	64%	60%	61,93%
Lesk Baseline EN-RU	47%	57%	51,51%
Lesk + Sent2Vec + Cosine Vectorial EN-RU	65%	61%	62,93%

Tables 3. Experimental Results for crosslingual dataset

We noticed to have lower scores when using the baseline (Lesk algorithm) than the combination of the three algorithms (Lesk, Sent2Vec and Cosine Vectorial).

Although it might have seemed to have similar results with the simple Word2Vec version, we considered Sent2Vec to be more reliable because it did not depend on the number of words in the vocabulary, and this was an advantage since the sizes of the vocabularies were diverse across languages.

5 Conclusion and Discussions

This paper presents a system participating at SemEval 2021 Task 2. Our solution indicates a good start for solving word sense disambiguation.

The main challenge behind word sense disambiguation is to make ample use of the available technologies since ambiguities in any language provide great difficulty in the use of information technology. The major difficulty lays in the fact that words in human language can be interpreted in more than one way, depending on the context (Tan, 2013).

Since we performed a detailed investigation of monolingual and bilingual disambiguation, our experimental results showed that Sent2vec and Lesk approaches are remarkably efficient for both tasks. The overall results are satisfactory and exceed the baseline; however, there is still room for improvement.

A larger and well-annotated dataset would provide more opportunities for exploring the issue of disambiguation. Additionally, building a dataset sufficient in size and diversity will allow experiments with deep learning methods. The biggest challenge in this project was working in different languages at the same time while some tools were available to English only.

From our research, we noticed that the average scores are around 75% when applying separately Lesk and Sent2Vec (or even Word2Vec), and it seems to be similar when combining the two.

References

Akyürek, A. F., Guo, L., Elanwar, R., Ishwar, P., Betke, M., Wijaya, D. T. (2020). *Multi-Label and Multilingual News Framing Analysis*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8614-8624).

Artetxe, M., Labaka, G. and Agirre, E. (2018). *A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings.* In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 789-798.

Bevilacqua, M., & Navigli, R. (2020). *Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2854-2864).

Baker, M., Francis, G., and Tognini-Bonelli, E. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*, Chapter 2. John Benjamins Publishing Company, Netherlands.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics.

Dinu, D., Lazaridou, A., and Baroni, M. (2015). *Improving zero-shot learning by mitigating the hubness problem*. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), workshop track.

Lefever, E., Hoste, V. (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 7th SemEval, pp. 158-166.

Daniel Loureiro and Al´ıpio 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.

Luan, Y., Hauer, B., Mou, L., Kondrak, G. (2020). *Improving Word Sense Disambiguation with Translations*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4055-4065.

Lu, A., Wang, W., Bansal, M., Gimpel, K., and Livescu, K. (2015). *Deep Multilingual Correlation for Improved Word Embeddings*. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 250-256.

Martelli F., Kalach N., Tola G., Navigli R. (2021) *SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC)*. In Proceedings of the 15th Workshop on Semantic Evaluation 2021.

Matusevych, Y. (2016). *Learning Constructions from Bilingual Exposure. Computational Studies of Argument Structure Acquisition*. PhD. Thesis, Tilburg University.

Merlo, P., Rodriguex, M. A. (2019). *Cross-lingual Word Embeddings and the Structure of the Human Bilingual Lexicon*. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 110-120.

Navigli, R. and Ponzetto, S. P. (2012). *Joining Forces Pays off: Multilingual Joint Word Sense Disambiguation*. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1399-1410.

Pagliardini Matteo, Gupta Prakhar, Jaggi Martin (2018) *Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features*. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)

Riley, P., Caswell, I., Freiag, M., Grangier, D. (2020). *Translationese as a Language in "Multilingual" NMT*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7737-7746.

Scarlini Bianca, Tommaso Pasini, and Roberto Navigli (2020) *SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation.* In Proceedings of the AAAI Conference on Artificial Intelligence.

Smith, S., L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). *Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax*. In Proceedings of 5th ICLR.

Tan, l. (2013). *Examining Crosslingual Word Sense Disambiguation*. MSc Thesis, Nanyang Technological University - http://compling.hss.ntu.edu.sg/pdf/2013-masters-tan-liling.pdf

Zhang, Y., Gaddy, D., Barzilay, R., and Jaakkola, T. (2016). *Ten pairs to tag – multilingual pos tagging via coarse mapping between embeddings*. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1307-1317.