PAW at SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation : Exploring Cross Lingual Transfer, Augmentations and Adversarial Training

Harsh Goyal, Aadarsh Singh and Priyanshu Kumar

Indian Institute of Technology (Indian School of Mines) Dhanbad, India
{harshgoyal13, aadarshsingh191198, kpriyanshu256}@gmail.com

Abstract

We experiment with XLM RoBERTa for Word in Context Disambiguation in the Multi Lingual and Cross Lingual setting so as to develop a single model having knowledge about both settings. We solve the problem as a binary classification problem and also experiment with data augmentation and adversarial training techniques. In addition, we also experiment with a 2-stage training technique. Our approaches prove to be beneficial for better performance and robustness.

1 Introduction

Language is complex even for human beings, let alone for computers. The same word serves different purposes in different scenarios, thus increasing the complexity of the Word Sense Disambiguation (WSD). For example in English, the word "bank" can refer to a financial institution or the land alongside a river. Many works revolving around WSD have been done with the help of explicit word sense inventories like WordNet [1] and BabelNet [2]. With the advent of advanced deep learning models, it is desirous to develop systems that have a good understanding of languages without such gold standards of word sense. This unsupervised learning can help the model learn better latent representations of words in different contexts.

In this paper, we aim to develop a single system that has knowledge of both multilingual and cross-lingual word sense disambiguation by training models with the combined data for both settings. We present our approaches for WSD in the multilingual and cross-lingual domain. The task is treated as a binary-classification problem: whether words have the same sense in the two given pairs of sentences. We experiment with XLM-RoBERTa (Conneau et al., 2019), which is based on the Transformer architecture (Vaswani et al., 2017), as the backbone of our architectures in both the settings. In addition, we also leverage external data and different training techniques and data augmentation.

The rest of the paper is organized as follows : various related works have been discussed in section 2, followed by a brief description of the shared task dataset in section 3. The system overview and experimental settings are covered in sections 4 and 5. Sections 6 contain the results. Section 7 concludes the paper and also includes scope of future work.

2 Related Work

Silberer and Ponzetto (2010) make use of graph algorithms for the word sense disambiguation task. They build a multilingual co-occurence graph in which the multilingual nodes are connected with translation edges and labelled with the target word's translations as obtained from the corresponding contexts.

Authors in Banea and Mihalcea (2011), use multilingual vector space which is obtained by expanding monolingual features engineered from more than one language, in order to generate a more effective, robust and utilitarian vector representation. These engineered features are then used for WSD.

Languages like Arabic do not have as many resources in the available dataset as compared to more common languages like English. To tackle this issue for the Persian language, Lefever and Hoste (2011) follow a two phase approach - in the first phase, they utilize an English Word sense disambiguation system to assign "sense tags" to words appearing in English sentences and then in the following phase, they transfer the senses obtained in the previous phase to corresponding Persian words.

In the Semeval-2013 WSD task (Navigli et al.,

[1] https://wordnet.princeton.edu/
[2] https://babelnet.org/

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 743–747
Bangkok, Thailand (online), August 5–6, 2021.

2013), Rudnick et al. (2013) take a classification-based approach to the Cross-Lingual WSD task. They build the HLTDI system in which they perform word alignment on the Europarl corpus. This helps them find samples in the training data which have ambiguous focus words. The paper describes three variants of the classifier - one is trained over local features, the second is trained over the data with translation of the focus word in the four target languages added to the feature vector and the final variant builds a Markov network of the first classifier in order to find the best translation.

A few works have also been submitted as a part of SemDeep-5 workshop (Espinosa-Anke et al., 2019). Ansell et al. (2019) make use of contextualised ELMo word embeddings. A Bidirectional Long Short Term Memory(LSTM) cell is used to extract better representation of the given sentences. To disambiguate the words, they optimise the cosine distance between the concatenated hidden representations of words, preceding and following the target word. Soler et al. (2019) augment the dataset by automatically substituting target words using contextual similarity. They then experiment with different contextual word embeddings and train a logistic regression classifier on top of that.

3 Dataset

The dataset (Martelli et al., 2021) [3] provided by the shared task organizers consists of both multilingual and cross-lingual data in English (EN), Arabic (AR), French (FR), Russian (RU) and Chinese (ZH). The dataset consists of two sentences and the words in corresponding sentences (which need disambiguation) and the corresponding label.

4 System Overview

Our experiments revolve around Facebook's XLM RoBERTa model, which was an update to their XLM-100 Language Model. XLM RoBERTa is based on the transformer architecture consisting of multi-attention heads which apply a sequence-to-sequence transformation on the input text sequence. The training procedure is inspired from RoBERTa (Liu et al., 2019) i.e. only the Masked Language Model objective is used. XLM RoBERTa is scaled up to 100 languages, thus becoming a good choice for multi-lingual datasets.

Another motivation to experiment with XLM RoBERTa comes from the facility of "Cross Lingual Transfer", which can help with unbalanced data of different languages. Knowledge is transferred for all languages if the model is trained for a particular task using data of only one language. Thus, this feature saves effort of gathering more data to make the data distribution balanced.

4.1 Problem Formulation

We perform experiments keeping the model architecture constant across all experiments. The model accepts both the sentences concatenated together. The input to the model is formulated as : $word_1 + < /s > + sentence_1 + < /s > + word_2 + < /s > + sentence_2$, where $< /s >$ is the separator token in XLM RoBERTa vocabulary.

Dropout is applied on the pooled encoding output from the model. The dropout probability is set to 0.3. The dropout applied output is then passed through a linear layer which provides us with the logits corresponding to the 2 classes.

4.2 Data Augmentation

Data augmentations are considered an important technique to avoid overfitting of neural networks thus making them more generalised. Since our model architecture accepts both the sentences together, there is room to apply a simple data augmentation during training. Consider t_1 and t_2 are the 2 sentences for a particular data instance. The training data is augmented as $t_1 \bigoplus t_2$ and $t_2 \bigoplus t_1$, where $\bigoplus$ represents concatenation. We apply the augmentation taking care that no data leakage takes place in the validation data.

4.3 Two Stage Training

To leverage the property of Cross Lingual Transfer, we first train the model on the WiC dataset (Pilehvar and Camacho-Collados, 2018), which consists only of English data. Then we train the same model (trained on WiC) on the MCL WiC dataset. This technique instills some knowledge via cross lingual transfer, about WSD in the first stage and then builds on the knowledge using the shared task dataset.

4.4 Adversarial Training (AT)

Adversarial training is another technique that is used to increase the robustness of models, which also helps in better generalisation. Adversarial training in Computer Vision is done by directly perturbing the input images. However, text data

[3] https://github.com/SapienzaNLP/mcl-wic

Model	EN-AR	EN-FR	EN-RU	EN-ZH
2 stage + Train Aug	75.3	78.2	78.7	75.2
2 stage + Train Aug + TTA	74.6	78.0	**78.9**	**75.3**
2 stage + Train Aug + AT	76.8	78.1	76.9	74.6
2 stage + Train Aug + AT + TTA	**76.9**	**79.5**	78.0	74.6
Best performance	89.1	89.1	89.4	91.2

Table 2: Test Scores for Cross Lingual Setting

being discrete is nature, the perturbations are added to the word embeddings.

Many approaches for adversarial training in NLP have been developed. We experiment using Miyato et al. (2016) approach with little modification. In their approach, the word embeddings are normalized first. Required perturbations are created using the gradients obtained via backpropagation. Let the sequence of (normalized) word embedding vectors of a text be $\boldsymbol{t}$. The model parameters are represented by $\boldsymbol{\theta}$. The probability of the text belonging to class y is given by $p(y|\boldsymbol{t};\boldsymbol{\theta})$. The adversarial perturbations $\boldsymbol{z}_{adv}$ are computed as follows:

$$\boldsymbol{g} = \nabla_t \, log \, p(y|\boldsymbol{t};\boldsymbol{\theta})$$

$$\boldsymbol{z}_{adv} = -\epsilon \boldsymbol{g} / \parallel \boldsymbol{g} \parallel_2$$

where ϵ is a hyper-parameter controlling the size of the perturbations. The adversarial loss is defined as :

$$\boldsymbol{L}_{adv}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} log \, p(y_n|\boldsymbol{t}_n + \boldsymbol{z}_{adv,n};\boldsymbol{\theta})$$

By using the gradients calculated from the above loss, the weights of the model are updated (the non-perturbed word embeddings of the model are updated). Our experiments deviate from the above method in the part that we do not normalize our pretrained word embedding of the model, since doing so might change the semantic meaning of the pretrained word embeddings. We perform adversarial training XLM RoBERTa model using $\epsilon = 1$.

Model	CV
W/O extra techniques	73.68
Train Aug	74.68
2 stage + Train Aug	75.64
2 stage + Train Aug + AT	**77.07**

Table 1: Cross Validation Scores

4.5 Test Time Augmentation (TTA)

The usage of the training data augmentation can be extended to test time as well. For a given data instance t1 and t2, the model predictions for $t_1 \bigoplus t_2$ and $t_2 \bigoplus t_1$ are combined using simple averaging of probabilities. Thus, this simple augmentation can help boost the performance of the model.

5 Experimental Setup

We make use of combined training and validation data provided by the shared task organizers. We perform a stratified 5 fold cross validation using the combined data. In all our experiments, we fine tune the entire model. Each fold is trained for 20 epochs using early stopping with patience of 6 and tolerance of 1e-3. The models are optimised using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 5e-6 and a batch size of 16 [4]. Inputs of maximum sequence length 172 are used in the model. The models have been implemented using Pytorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2019) library.

6 Results

Accuracy score is the official evaluation metric for the shared task. The test predictions are obtained by combining the predictions of all the 5 fold models (by averaging the predictions from all models). Table 1 lists down the cross validation accuracy scores of all the experiments. The test scores are categorised as cross-lingual and multilingual and are presented in tables and 2 and 3 respectively . For bench marking purpose, we also mention the best performances achieved by participants of the shared task.

A few observations can be made by looking at the results:

[4]It is important to note that XLM RoBERTa requires a much smaller learning rate as compared to BERT and other models, for training; XLM RoBERTa is incapable of learning if trained using high learning rates like 2e-5

Model	EN-EN	AR-AR	FR-FR	RU-RU	ZH-ZH
2 stage + Train Aug	84.5	79.7	78.8	77.4	79.6
2 stage + Train Aug + TTA	**85.4**	**80.2**	79.0	**77.7**	**80.4**
2 stage + Train Aug + AT	85.2	79.0	80.2	76.9	79.2
2 stage + Train Aug + AT + TTA	85.1	80.0	**80.5**	77.6	79.7
Best performance	93.3	84.8	87.5	87.4	91.0

Table 3: Test Scores for Multilingual Setting

1. Test Time Augmentation helps in boosting the scores.

2. In the cross lingual setting, models trained with and without adversarial training are competent to the same extent. On the other hand, in the multilingual setting, models trained without adversarial training seem to have the upper hand.

7 Conclusion and Future Work

We explore the performance of XLM RoBERTa at Word In Context Disambiguation both in the multilingual and cross lingual setting. We also explore different training techniques such as two-stage training and adversarial training along with some simple augmentations to make our models robust and more generalized. Test Time Augmentations, based on training augmentation turn out to useful. For future work, we can explore the performance of ensembling different kinds of models trained with and without adversarial training together, so as to produce more robust results. It will also be interesting to experiment with larger backbone models in the current architecture.

Acknowledgments

We thank Kaggle for providing free GPU and TPU services.

References

Alan Ansell, Felipe Bravo-Marquez, and Bernhard Pfahringer. 2019. An ELMo-inspired approach to semdeep-5's word-in-context task. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 21–25.

Carmen Banea and Rada Mihalcea. 2011. Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Luis Espinosa-Anke, Thierry Declerck, Dagmar Gromann, Jose Camacho-Collados, and Mohammad Taher Pilehvar, editors. 2019. *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*. Association for Computational Linguistics, Macau, China.

Els Lefever and Veronique Hoste. 2011. Examining the validity of cross-lingual word sense disambiguation. *Polibits*, (43):29–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

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)*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca

Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037.

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*.

Alex Rudnick, Can Liu, and Michael Gasser. 2013. Hltdi: Cl-wsd using markov random fields for semeval-2013 task 10. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 171–177.

Carina Silberer and Simone Paolo Ponzetto. 2010. UHD: Cross-lingual word sense disambiguation using multilingual co-occurrence graphs. In *Erk K, Strapparava C, editors. Proceedings of the 5th International Workshop on Semantic Evaluation; 2010 Jul 15-16; Uppsala, Sweden. Stroudsburg: ACL; 2010. p. 134-7.* ACL (Association for Computational Linguistics).

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's Transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.