Validation of Language Agnostic Models for Discourse Marker Detection

Mariana Damova[⋆]
[⋆]**Mozaika, Ltd.**
[⋆]mariana.damova@mozajka.co

Giedrė Valūnaitė Oleškevičienė[⊙]
[⊙]**Mykolas Romeris University**
[⊙]gvalunaite@mruni.eu

Purificação Silvano[†]
[†] **Centre of Linguistics of the University of Porto**
[†]msilvano@letras.up.pt

Ciprian-Octavian Truică[§,‡]
[§]**Uppsala University**
[§]ciprian-octavian.truica@it.uu.se

Christian Chiarcos[§,‡]
Goethe-Universität
chiarcos@cs.uni-frankfurt.de

Kostadin Mishev[◦]
[◦]**Ss. Cyril and Methodius University**
[◦]kostadin.mishev@finki.ukim.mk

Chaya Liebeskind[⋄]
[⋄]**Jerusalem College of Technology**
[⋄]liebchaya@gmail.com

Dimitar Trajanov[◦]
[◦]**Ss. Cyril and Methodius University**
[◦]dimitar.trajanov@finki.ukim.mk

Elena-Simona Apostol[§,‡]
[‡]**University Politehnica of Bucharest**
[‡]elena.apostol@upb.ro

Anna Bączkowska[×]
[×]**University of Gdansk**
[×]anna.baczkowska@ug.edu.pl

Abstract

Using language models to detect or predict the presence of language phenomena in the text has become a mainstream research topic. With the rise of generative models, experiments using deep learning and transformer models trigger intense interest. Aspects like precision of predictions, portability to other languages or phenomena, scale have been central to the research community. Discourse markers, as language phenomena, perform important functions, such as signposting, signalling, and rephrasing, by facilitating discourse organization. Our paper is about discourse markers detection, a complex task as it pertains to a language phenomenon manifested by expressions that can occur as content words in some contexts and as discourse markers in others. We have adopted language agnostic model trained in English to predict the discourse marker presence in texts in 8 other unseen by the model languages with the goal to evaluate how well the model performs in different structure and lexical properties languages. We report on the process of evaluation and validation of the model's performance across European Portuguese, Hebrew, German, Polish, Romanian, Bulgarian, Macedonian, and Lithuanian and about the results of this validation. This research is a key step towards multilingual language processing.

1 Introduction

Using language models to detect or predict the presence of language phenomena in the text has become a mainstream research topic. The performance of these models heavily depends on the quantity and on the quality of the data used for training them. Producing datasets of training data is a very time-consuming and expensive process, requiring human expertise. Deep learning models have been so far built by training single languages one by one. This requires the availability of training data in each language of interest, and makes obtaining language models for multiple languages complicated, expensive and virtually impossible for smaller or rare languages. That is why research efforts have been focusing on removing the need for manual preparation of training data by developing deep learning architectures able to produce language models for languages without training on them - language agnostic models. Language agnostic models build models based on training data in one language, and then extrapolate them to other unknown for the model languages. It is important to know how well they perform and whether the quality of the prediction results in unseen languages is good enough to adopt and further develop these approaches and architectures. This paper presents experiments with a language-agnostic model in 8 languages, trained on data in English, to detect the presence and absence of discourse markers in unseen text and discusses the process and the results of validating their performance, demonstrating the good performance and the viability of the model. In our case, the model targets discourse markers, essential pointers for the communicational setting and the speaker's attitudes. They have particular roles in facilitating discourse organization and providing text coherence and cohesion between discourse segments.

The structure of the paper is as follows: Section 2 presents related work; Section 3 describes the language-agnostic machine learning method that has been adopted for the experiment; Section 4 gives an overview of the multilingual corpus used in the experiment; Section 5 describes the experiment, discusses the validation process and the performance of the language-agnostic model; Section 6 concludes the paper.

2 Related work

Regarding NLP tasks, there have been advancements in identifying and classifying discourse markers. For instance, Zufferey (2004) describes an experiment where discourse markers are detected and assigned inferential semantic functions. For the improvement of automatic methods for discourse markers detection and classification, shared tasks such as DISRPT 2019 and 2021 editions (Zeldes et al., 2019, 2021) and Discourse Relation Classification across RST (Mann and Thompson, 1988), SDRT (Asher et al., 2003), and PDTB (Prasad et al., 2008) have played a significant role. Following CoNLL 2015 setting, Kurfali (2020) developed an experiment to determine the efficacy of pre-trained language models in the task of shallow discourse parsing (SDP) used to identify explicit local discourse relations without resorting to tree/ graphs structures. The BERT-based model and the Hugging's face Transformer library were employed with the maximum sequence length 400 for the first approach and 250 for the second. For the test set, the author used PDTB. The model evaluation was performed on top of the official results of CoNLL 2015 (Xue et al., 2015) and 2016 (Xue et al., 2016) shared tasks, and of (Knaebel et al., 2019). Regarding connective identification, the model accomplished an F1-score of 95.76%, similar to previous experiments. In the 2021 edition of the DISRPT Shared Task, the system with the best results was DisCoDisCo (Gessler et al., 2021) with a Transformer-based neural classifier. This model outperformed state-of-the-art scores from the 2019 DISRPT concerning connective detection with an F1-score of 91.22%.

3 Language agnostic methods

Language-agnostic models have been developed to allow cross-language analysis and language phenomena detection without the need to process training data in each language manually. Such model is La-BSE, which we have adopted for our experiment, based on the amount of languages it is able to cover and on its modeling architecture.

The Google's language-agnostic BERT sentence embedding (La-BSE) model supports 109 languages (Feng et al., 2020). The multilingual architecture of BERT is adapted to produce language-agnostic sentence embeddings for 109 languages. La-BSE combines the masked-language model (MLM) and translation language model (TLM) pretraining with a translation ranking task using bidirectional dual encoders. This method improves the average bi-text retrieval accuracy and establishes new state-of-the-art on the bi-text retrieval.

4 Datasets

The multilingual datasets that have been part of the experiment contain examples from nine languages English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Polish and European Portuguese, compiled from the publicly available TED Talk transcripts. It is an ongoing expansion of TED-EHL parallel corpus LINDAT/CLARIN-LT repository [1]. In addition, we have produced a list

[1] http://hdl.handle.net/20.500.11821/34

of multiword expressions (MWE) that can occur as discourse markers in specific contexts and as content expressions in others, where ambiguity is tricky to capture. For example, the expression *you know* in examples 1 and 3 below describes the content, whereas in example 2 it describes a discourse marker.

1. By the way, just so *you know*
2. But *you know*, they have, after all, evolved in a country without telephones,
3. *you know* what I mean.

Expressions of this nature are also *I remember*, *I mean*, *I think*, *you see*, etc.

Other MWEs from the established discourse markers list are lexicalized discourse markers that are interpreted as such in any context. Such MWEs are *of course*, *for example*, *above all*, *in addition* and the like.

We have produced eight bilingual datasets with aligned parallel texts in English and another language, based on the occurrece of MWE potentially describing a discourse marker in the sentence context. The structure of English part of the the aligned bilingual corpus is shown in table 1.

In the bilingual parallel corpus, another four columns to the right of the last column of the data for English contain the translations of the English examples in the given language from the eight we cover. So, we end up with a corpus of eight bilingual parallel aligned corpora with an overall size presented in table 2.

5 Experiment

The English dataset was used as a baseline. It is composed of the union of all unique sentence contexts from all language pairs, and counts 44,209 sentence contexts. From them 4777 have been manually annotated, and 1019 turned to be with a discourse marker present (1) whereas 3758 - without a discourse marker present (0). The English dataset was split 80% for training and 20% for testing. The training set is used to fine-tune the XLM-RoBERTA Large model for the classification. The test set is used to evaluate the performance on unseen samples to predict the presence or absence of discourse markers in the training dataset.

The same training dataset was used to train with the La-BSE language-agnostic method to generate a model that has been consequently run through all languages from the bilingual parallel corpus (cf. table 2 described above). As a result, prediction for the presence or absence of discourse markers in each context for each language has been generated and output in the table structure shown in table 3. Note that the English example does not have a value for presence or absence of a discourse marker in the context (9) in table 3. This indicates that the trained model in English has been run through unseen examples in the other languages.

6 Validation

The validation of the results has had two stages. In the first stage, the prediction results have been verified against the manual annotations. Table 4 shows the evaluation for Bulgarian and Lithuanian with considerably better prediction results for Lithuanian - 0.94 precision than for Bulgarian - 0.74 precision.

As a second step, human experts manually validated the predictions of the language-agnostic model. To provide the most accurate possible outlook, we took the first 100 lines of each bilingual file, ensuring that all selected examples differ.

Then, human experts had to evaluate whether the prediction of the model was correct or not. The validation has shown that the La-BSE method, trained on English text, performs very well on unseen languages regardless of their family and on diverse unseen texts. The results are shown in table 5 below with an average of 12 wrongly predicted occurrences and 88% precision.

The reasons for the discrepancies in the correct prediction rate are still to be analyzed. We predict that they may be related to the texts themselves, the human analysts' expert judgement, and the structure of the language compared to the structure of English.

7 Conclusion

This paper presented an experiment of applying a language-agnostic machine learning method to a multilingual corpus of 9 languages to verify how well it would perform detecting discourse markers when trained in English. The two validation methods with testing corpus and with human expert assessment showed only a little discrepancy in the analysis of the results. The human expert analysis performed better than the automatic evaluation of the testing corpus. The reasons for these discrepancies are to be investigated in detail in our future

Table 1: Structure of the English part of the corpus

MWE	Sentence chunk	Context	Discourse Marker Presence
I remember	And I remembered that the old and drunken guy destroying my statistical significance of the test. So I looked carefully at this guy. He was 20-some years older than anybody else in the sample.	And I remembered that the old and drunken guy came one day to the lab wanting to make some easy cash	0
You know	But you know, these stories, because he would have pulled the mean of the group lower, giving us even stronger statistical results than we could. So we decided not to throw the guy out and to rerun the experiment.	But you know, these stories, and lots of other experiments that we've done on conflicts of interest, basically kind of bring two points	1

Table 2: Constituted multilingual datasets

language	aligned sentences with MWE
English	43600
Macedonian-English	2846
German-English	15852
Lithuanian-English	4112
Bulgarian-English	19209
Portuguese-English	4398
Polish-English	17408
Romanian-English	18946
Hebrew-English	23566

Table 3: Example of model output

DM EN	S Chunk EN	DM Presence EN	text LANG	LABSE prediction
in fact	In fact, she had aged a lotṪhe woman who as a child had skipped with him through fields and broken his heart	9	Всъщност, доста беше остаряла Жената, която като дете бе подскачала с него през полята и бе разбила сърцето му	1

work. This experiment proved that the language-agnostic models' performance is not affected significantly by the structure of the language or other lexical or grammatical peculiarities of the single languages and gives a good prediction for the presence of discourse markers in texts in unseen by the model languages.

Table 4: Language-Agnostic Methods Results

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
La-BSE (BG)	0.7273	0.7403	0.7090	0.7459	0.7243	0.4551
La-BSE (LT)	0.8338	0.9412	0.8758	0.2877	0.9073	0.1228

Table 5: Human validation results

Language	Number of Wrong Predictions	Total Number of Examples	Precision ratio
BG	10	100	0,90
MK	19	100	0,81
EN	16	100	0,84
HE	5	100	0,95
PT	20	100	0,80
DE	17	100	0,83
PL	10	100	0,90
LT	12	100	0,88
RO	1	100	0,99

Acknowledgements

This work has been done within the COST Action CA18209 - NexusLinguarum.

References

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

René Knaebel, Manfred Stede, and Sebastian Stober. 2019. Window-based neural tagging for shallow discourse argument labeling. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 768–777, Hong Kong, China. Association for Computational Linguistics.

Murathan Kurfali. 2020. Labeling explicit discourse relations using pre-trained language models. *ArXiv*, abs/2006.11852.

William Mann and Sandra Thompson. 1988. Rethorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. Introduction to discourse relation parsing and treebanking (DISRPT): 7th workshop on Rhetorical Structure Theory and related formalisms. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 1–6, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandrine Zufferey. 2004. Une analyse des connecteurs pragmatiques fondée sur la théorie de la pertinence et son application au TALN. *Nouveaux Cahiers de Linguistique Française*, 25:257–272.