Workshop on

Deep Learning and Neural Approaches for Linguistic Data

Skopje, North Macedonia & online
30 September 2021

Book of abstracts

Scientific committee:

Eliot Bytyçi, University of Prishtina
Radovan Garabík, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Dagmar Gromann, University of Vienna
Chaya Liebeskind, Jerusalem College of Technology, Lev Academic Center
Giedrė Valūnaitė Oleškevičienė, Mykolas Romeris University
Hugo Gonçalo Oliveira, University of Coimbra
Purificação Silvano, University of Porto

Editor: Radovan Garabík
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences

Nexus Linguarum

https://nexuslinguarum.eu/

This publication is based upon work from COST Action CA18209 – European network for Web-centred linguistic data science, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

www.cost.eu

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

COST is supported by the Horizon 2020 Framework Programme of the European Union

Foreword

Deep learning and neural approaches are indispensable in modern Natural Language Processing and generally in all kinds of linguistic data analysis tasks. This workshop is aimed at deep learning in connection with linguistic data and the effective use of deep learning in understanding the specificities of linguistic data. The submissions collected in this book of abstracts deal with deep learning used to improve named entity recognition; BERT in conjunction with a compilation of lexical patterns to automatically acquire lexico-semantic relations; using transformer models to predict discourse relations and speaker's attitudes; using transformer models to automatically extract terminological concept systems; and an automatic detection of rhetorical patterns in academic texts using machine learning algorithms designed for image object detection purposes trained on the page layout and graphical elements.

The workshop shows just a small fraction of the variety of problems that modern deep learning methods can successfully tackle, and demonstrates the usefulness of linguistic linked open data, as results of and interconnected with neural approaches.

Radovan Garabík, Dagmar Gromann

Table of Contents

Speaker Attitudes Detection through Discourse Markers Analysis

Giedrė Valūnaitė Oleškevičienė[1], Chaya Liebeskind[2], Dimitar Trajanov[3], Purificação Silvano[4], Christian Chiarcos[5], Mariana Damova[6]

[1]Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania
gentrygiedre@gmail.com
[2]Jerusalem College of Technology, Jerusalem, Israel
liebchaya@gmail.com
[3]Ss. Cyril and Methodius University – Skopje, North Macedonia
dimitar.trajanov@gmail.com
[4]University of Porto, Portugal
puri.msilvano@gmail.com
[5]University of Frankfurt, Frunkfurt, Germany
chiarcos@informatik.uni-frankfurt.de
[6]Mozaika Ltd, Solunska 52, Sofia 1000, Bulgaria
mariana.damova@mozajka.co

Keywords: discourse markers, speaker attitudes detection, annotation, linguistic linked open data, transformer models, machine learning.

Extended Abstract

Speaker attitude detection is important for processing opinionated text. Survey data as such provide a valuable source of information and research for different scientific disciplines. They are also of interest to practitioners such as policymakers, politicians, government bodies, educators, journalists, and all other stakeholders with occupations related to people and society. Survey data provide evidence about particular language phenomena and public attitudes to provide a broader picture about the clusters of social attitudes. In this regard, attitudinal discourse markers play a central role in the sense that they are pointers to the speaker's attitudes. These single word or multiword expressions (MWE) are mainly drawn from syntactic classes of conjunctions, adverbials, and prepositional phrases (Fraser, 2009), as well as expressions such as *you know*, *you see*, and *I mean* (Schiffrin, 2001; Hasselgren, 2002; Maschler & Schiffrin, 2015). Discourse markers are regarded as significant discourse relations' triggers, and, consequently, are largely studied (e.g. Sanders et al. 1992; Knott & Dale 1994; Wellner et al 2006; Taboada & Das 2013; Das 2014; Das & Taboada 2019; Silvano 2011). Recently, discourse relations and discourse marker research has gained certain impetus with corpora annotation for exploring discourse structure in texts, for example, RST-DT English corpus (Carlson, Marcu & Okurowski 2003); Penn Discourse Tree Bank (PDTB) (Prasad et al. 2008); SDRT Annodis French corpus (Afantenos et al., 2012).

The large bulk of these corpora is manually annotated, mostly by trained linguists, less by non-experts, and only a reduced number undergoes automatic/semiautomatic annotation (with human supervision).

This study describes ongoing work whose ultimate goals are: (i) to collect methods for appropriate processing of free text answers to open questions in surveys with respect to speaker attitudes identified by discourse markers; and (ii) to establish guidelines for the creation of LLOD vocabularies for discourse markers. In particular, this paper presents the process of constituting a multilingual corpus, creating an annotation schema of discourse relations for marking the discourse markers, and applying machine learning transformer models to predict their appearance in unknown texts. We apply a two-step approach to detecting speaker attitudes by identifying discourse markers and the semantics of the discourse relations they introduce in text using neural machine learning transformer models to ensure the interlinking of multilingual discourse markers.

To achieve the aforementioned goals, so far, we have created a parallel corpus containing data from 6 languages, using the publicly available TED Talk transcripts. It is an ongoing expansion of TED-EHL parallel corpus published in LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34. The multilingual corpus contains alignments of Lithuanian, Bulgarian, Portuguese, Macedonian, and German languages with English as pivot language with a size of 1.3 million sentences. Secondly, we constitute a vocabulary of multiword expression that can play the role of discourse markers in text based on theoretical insights by Schiffrin (1987) and classification provided by Fraser (2009). The next step was the manual annotation of the 2428 English-Bulgarian-Lithuanian aligned sentences containing the multiword expressions (MWE) as discourse markers or content expressions (1 or 0). Example (1) below classifies the multiword expression *you know* as a discourse marker (annotated 1) used to introduce a new discourse message, whereas example (2) represents content words (annotated 0) fully integrated into the sentence.

(1) That's ridiculous. *You know*, this is New York, this chair will be empty, nobody has time to sit in front of you.
(2) *You know* some people who say “Well”

The annotated corpora have been used to train machine learning models to predict the existence of discourse markers in a text. Because we had a multilingual dataset, we chose FastText (Joulin et al. 2016) XLM-Roberta (Conneau et al. l 2019) as the base models. The model was fine-tuned using the ktrain library (Maiya 2020), a low-code Python library built on top of the state-of-the-art Transformers library (Wolf et al. 2020). The dataset was divided 80-20 for train and test datasets, and the model was trained using a learning rate of 0.00001 for three epochs. The dataset was slightly unbalanced (53% records without a discourse marker and 47% with a discourse marker), so we used class balancing weights to compensate. The model fine-tuning was run ten times, and the average performance is reported in Table 2.

Table 1 shows an example of annotated corpus used for training the transformer models.

Table 1: Example of annotated corpus entries

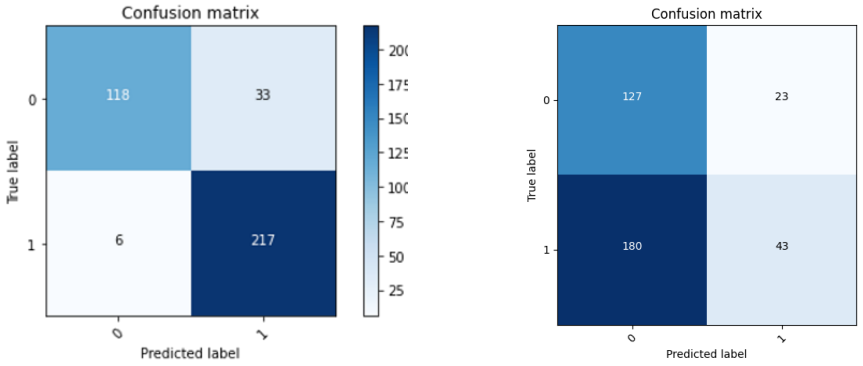
MWE	Sentence chunk	Context	Discourse Marker Presence
I remember	And I remembered that the old and drunken guy	destroying my statistical significance of the test. So I looked carefully at this guy. He was 20-some years older than anybody else in the sample. And I remembered that the old and drunken guy 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

The results of the two trained models for English is given in table 2 and figure 1 below. As this is the first attempt to identify the presence of discourse markers in unseen text with transformer models we think the results are promising.

Table 2: Results FastText XLM-RoBERTa-Large

	FastText	XML-Roberta-Large
Accuracy	0.46	0.90
Precision	0.65	0.87
Recall	0.19	0.97
Specificity	0.85	0.78
F1-Score	0.30	0.90
MCC	0.05	0.79

Figure 1: Confusion matrices – FastText and XLM-RoBERTa-Large



Regarding the semantics of discourse markers, we are adopting ISO 24618-8 annotation scheme to semantically annotate discourse relations as carriers of speaker attitudes in English, and Chiarcos (2014) methodology to represent them as LLOD and extend the semantic vocabularies of discourse relations (reference). Consequently, we will apply transformer models to predict the semantics of present discourse markers in unseen text in the 6 languages of the research.

References

[1] Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolle, J., Tanguy, L., Vergez-Couret, M. & Vieu, L. (2012) An empirical resource for discovering cognitive principles of discourse organization: The ANNODIS corpus. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), Proceedings of the 8th International Conference on Language Resources and Evaluation – LREC 2012 (pp. 2727-2734). Luxembourg: European Language Resources Association.

[2] Carlson, L.; Marcu, D. & Okurowsi, M. E. (2003) Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the second Sigdial Workshop on discourse and dialogue*. https://aclanthology.org/W01-1605

[3] Christian Chiarcos (2014). Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).*
http://www.lrec-conf.org/proceedings/lrec2014/pdf/893_Paper.pdf

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[5] Das, D. & Taboada, M. (2019) Multiple Signals of Coherence Relations. *Discourse* [Online], 24: 1-38.

[6] Das, D. (2014) *Signalling of Coherence Relations in Discourse.* PhD dissertation. Simon Fraser University, Burnaby, Canada.
https://summit.sfu.ca/item/14446

[7] Fraser, B. (2009) An account of discourse markers, International review of Pragmatics 1(2), 293–320, Publisher: Brill

[8] Hasselgren, A. (2002) Learner corpora and language testing: Small words as markers of learner fluency, Computer learner corpora, second language acquisition and foreign language teaching, 143–174, Publisher: John Benjamins Amsterdam, The Netherlands.

[9] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jegou, H., Mikolov, T. (2016) Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651

[10] Knott, A. and R. Dale (1994) "Using Linguistic Phenomena to Motivate a Set of Coherence Relations", in *Discourse Processes*, 18, (1), 35-62.

[11] LINDAT/CLARIN-LT repository http://hdl.handle.net/20.500.11821/34 (2021)

[12] Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. arXiv preprint arXiv:2004.10703. 11. Maschler, Y. and Schiffrin, D. (2015) Discourse markers: Language, meaning, and context, The handbook of discourse analysis, 1, 189-221. Publisher: Wiley Online Library.

[13] Prasad, R.; Dinesh, N.; Lee, A.; Miltsakaki, E.; Robaldo, L.; Joshi, A. and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008)*. https://aclanthology.org/L08-1093/

[14] Sanders, T.; J. W. Spooren and Leo G. M. Noordman (1992) Toward a Taxonomy of Coherence Relations. Discourse Processes, 15, 1-35.

[15] Schiffrin, D. (2001) Discourse markers: Language, meaning, and context, The handbook of discourse analysis 1, 54–75, Publisher: Wiley Online Library.

[16] Silvano, P. (2011) Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese. PhD Thesis, University of Porto.
https://repositorio-aberto.up.pt/handle/10216/56024?locale=pt

[17] Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).