Workshop on

# Deep Learning and Neural Approaches for Linguistic Data

Skopje, North Macedonia & online
30 September 2021

Book of abstracts

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

# Multilingual Extraction of Terminological Concept Systems

Dagmar Gromann[1][0000−0003−0929−6103], Lennart Wachowiak[1], Christian Lang[1], and Barbara Heinisch[1]

University of Vienna, Gymnasiumstraße 50, 1190 Vienna, Austria
{dagmar.gromann,lennart.wachowiak,christian.lang,barbara.heinisch}@univie.ac.at
https://transvienna.univie.ac.at/

## Extended Abstract

Terminological Concept Systems (TCS) provide a means of organizing, structuring and representing domain-specific multilingual information and are important to ensure terminological consistency in many tasks, such as translation and cross-border communication. Several methods for Automated Term Extraction (ATE) have been proposed that extract terms, i.e., single- or multi-word sequences, from domain-specific texts. ATE plays a role in many NLP tasks, such as information extraction, knowledge graph learning, and text summarization. An initial classification of ATE methods into statistical, linguistic or hybrid has recently been refined by [1] to methods based on term occurrence frequencies (e.g. C/NC-value), occurrence contexts, domain-specific corpora combined with general language corpora (e.g. Weirdness), topic modeling, and those utilizing Wikipedia. Even though the use of neural networks in ATE is mostly limited to generating embeddings, few exceptions exist that could not be accommodated by this classification.

A first use of BERT-based language models is documented by [3] and [9] rely on LSTM, GRU and BERT embeddings to achieve high F1 scores for Lithuanian ATE in the cybersecurity domain. Inspired by this first success of transformer-based models, we investigated two variations of the multilingual pretrained language model XLM-RoBERTa (XLM-R) [2] with an innovative use of the multilingual pretrained NMT model mBART [6]. Taking a natural language sentence as input, the model should predict all sequences of varying length that represent a domain-specific term. For instance, for the sentence "We meta-analyzed mortality using random-effect models" the model should output the individual terms *meta-analyzed*, *mortality* and *random-effect models*. Our best model, an XLM-R fine-tuned sequence classifier [5], outperformed the BERT-based baselines by 9 to almost 12% F1 score in English, French, and Dutch of the ACTER [8] and performed well for the ACL RD-TEC 2.0 dataset [7] without baseline.

In order to extract a TCS, a method to detect interrelations between extracted terms across languages is required. To the best of our knowledge this combination has not been proposed and relation extraction focuses mostly on extracting named entities and their interrelations. We present a method and tool called Text2TCS[1] that automatically extracts terms (including named

[1] https://text2tcs.univie.ac.at/

entities), groups them by synonymy into concepts, and detects their interrelations from text. We consider a pre-specified typology of terminological relations common in terminology science from hierarchical, i.e., generic and partitive, to non-hierarchical, e.g. ownership, instrumental, spatial relations. For instance, from the example terms previously extracted the model should predict an instrumental relation going from *random-effect models* to *meta-analyzed*.

The objective to extract a TCS from text in one language with a pre-specified relation typology is to facilitate the comparison to a TCS extracted from a text in another language. Since relations not only exist between terms that occur in the same sentence, we trained an intra-sentence level model [11] and complemented it with a document-level relation extraction model that is able to detect term relations without context and irrespective of the position of the terms in text (this model is based on our winner of the CogALex-VI Shared Task [10]). For the joint relation and term extraction we tested several existing datasets, appropriating them to the typology of relations we utilize, especially SemEval 2010 Task 8 [4]. However, the distribution of relations in those datasets is biased towards generic relations, which is one of the reasons why we decided to provide our own gold standard annotations in German and English by two terminological experts and a silver standard annotation in several languages by students of a translation master. We consider the latter a silver standard, since students were asked to provide the data in German and one or two other languages of their choice depending on availability. The comparison to the German gold standard showed a lower number of annotations in student works.

A TCS serves the objective to explicitly structure terminological knowledge and relations implicit in a text and thereby aid specialized communication and knowledge transfer. Training based on fine-tuned pre-trained Transformer models has focused on English and German, evaluation was additionally performed on Spanish, Portuguese, French, Italian, Romanian, and Russian, and in total supports at least 22 languages at inference time[2]. The tool will soon be available on the European Language Grid[3]. A TCS is an important language technology to generate language resources for the Linguistic Linked Open Data (LLOD) cloud. Currently, Text2TCS outputs TBX/XML as well as a TSV-based generic format and we intend to complement TBX/RDF to facilitate its LLOD-compatibility.

## Bibliography

[1] Astrakhantsev, N.: ATR4S: toolkit with state-of-the-art automatic terms recognition methods in scala. Language Resources and Evaluation **52**(3), 853–872 (2018)

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguis-

[2] See https://text2tcs.univie.ac.at/text2tcs-dokumentation/ for a complete list

[3] https://live.european-language-grid.eu/catalogue/tool-service/1315

tics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://www.aclweb.org/anthology/2020.acl-main.747

[3] Hazem, A., Bouhandi, M., Boudin, F., Daille, B.: TermEval 2020: TALN-LS2N system for automatic term extraction. In: Daille, B., Kageura, K., Terryn, A.R. (eds.) Proceedings of the 6th International Workshop on Computational Terminology. pp. 95–100. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.computerm-1.13

[4] Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In: Erk, K., Strapparava, C. (eds.) Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://www.aclweb.org/anthology/S10-1006

[5] Lang, C., Wachowiak, L., Heinisch, B., Gromann, D.: Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3607–3620. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.316, https://aclanthology.org/2021.findings-acl.316

[6] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics **8**, 726–742 (2020). https://doi.org/10.1162/tacl_a_00343

[7] QasemiZadeh, B., Schumann, A.K.: The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1862–1868. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://www.aclweb.org/anthology/L16-1294

[8] Rigouts Terryn, A., Hoste, V., Lefever, E.: In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. Language Resources and Evaluation **54**, 385–418 (2019). https://doi.org/10.1007/s10579-019-09453-9

[9] Rokas, A., Rackevičienė, S., Utka, A.: Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In: Human Language Technologies–The Baltic Perspective. vol. 328, pp. 39–46. IOS Press (2020). https://doi.org/10.3233/FAIA200600

[10] Wachowiak, L., Lang, C., Heinisch, B., Gromann, D.: CogALex-VI shared task: Transrelation - a robust multilingual language model for multilingual relation identification. In: Xiang, R., Chersoni, E., Iacoponi, L., Santus, E. (eds.) Proceedings of the Workshop on the Cognitive Aspects of the Lexicon. pp. 59–64. Association for Computational Linguistics, Online (Dec 2020), https://www.aclweb.org/anthology/2020.cogalex-1.7

[11] Wachowiak, L., Lang, C., Heinisch, B., Gromann, D.: Towards learning terminological concept systems from multilingual natural language text. In: 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik (2021)

# 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 k-train 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*

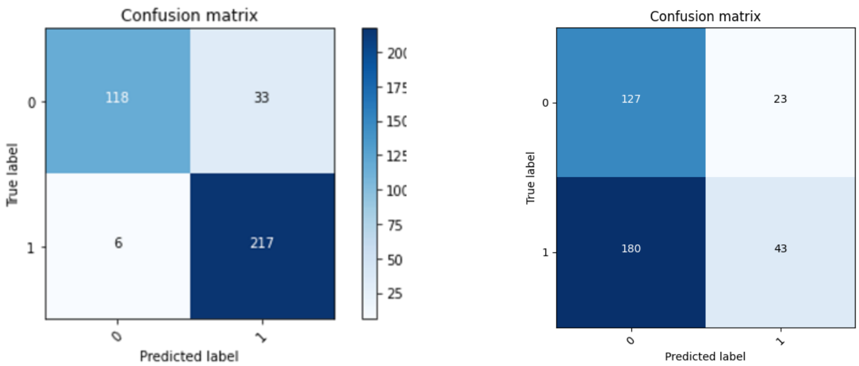
| MWE | Sentence chunk | Context | Discourse Marker Presence |
|---|---|---|---|
| I remember | And I remembered that the old and drunken guy | destroying my statistical significance of the test. So I looked carefully at this guy. He was 20-some years older than anybody else in the sample. And I remembered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

# Acquiring Lexico-Semantic Knowledge from a Portuguese Masked Language Model*

Hugo Gonçalo Oliveira

CISUC, Department of Informatics Engineering
University of Coimbra, Coimbra, Portugal
hroliv@dei.uc.pt

## Extended Abstract

Whether for the creation or for the enrichment of lexical knowledge bases (LKBs) like WordNet [4], there is a long research history on the automatic acquisition of lexico-semantic knowledge from textual corpora. Following the seminal work of Hearst [9], much relies on lexico-syntactic patterns where related words tend to occur, e.g., "$X_1$`, such as` $X_2$", for $X_1$ is-a-hypernym-of $X_2$.

In the last decade, research interest shifted to efficient models of distributional semantics, where words are represented by vectors learned from large corpora, also a friendlier format for machine learning. Word2vec [10] or GloVe [12] are good for computing word similarities, but fail to have explicit representations of semantic relations, even if some can be obtained through analogy [5].

More recently, transformer-based models, like BERT [2] or GPT [14], became the paradigm. They are useful for a broad range of tasks, but are also not ready for providing explicit semantic relations. Yet, they provide a shortcut for earlier corpora-based approaches, because they are pre-trained in large collections of text and are good at filling blanks or computing the probability of sentences, including those using the aforementioned patterns. It is thus no surprise that such models have been assessed for the presence of relational knowledge [13], for relation induction [1], and it has been noted that they perform particularly well in the acquisition of hypernyms [3]. While the previous target English, recent work [11] has exploited BERT for detecting hyponymy pairs in Portuguese.

In this work, we explore BERTimbau [15] base, a BERT model pre-trained for Portuguese, in the acquisition of lexico-semantic relations. For this, we compiled a list of patterns for the relations covered by TALES [7], a dataset created for assessing lexico-semantic analogies in Portuguese. For each relation, TALES includes 50 entries with two columns: a word (question) and a list of related words (answers). An example for hyponymy-of is: `água líquido/substância` (water liquid/substance). Some considerations had to be made when creating the patterns, such as avoiding patterns starting with a mask, because many suggested fillers were functional words; or including patterns for both masculine and feminine arguments.

* This work was partially supported by: the COST Action CA18209 Nexus Linguarum (European network for Web-centred linguistic data science); national funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project CISUC – UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020.

Each pattern was used to predict the answers given the question words in TALES, and their accuracy was compared with LRCos [5] computed in a GloVe model for Portuguese [8][1]. The latter was outperformed for six relations (out of 14), for which table 1 presents the best-performing pattern, their accuracy and accuracy at the top-10 answers, in comparison with LRCos. For each entry of the target type in TALES: $X_1$ was replaced by the question word and predictions for the `[MASK]` tag were used as answers.

| Relation | BERT | | | LRCos | |
|---|---|---|---|---|---|
| | **Pattern** | **Acc** | **Acc@10** | **Acc** | **Acc@10** |
| Antonym-of | `ou [MASK] ou` $X_1$ | 0.32 | 0.42 | 0.20 | 0.46 |
| Hypernym-of (abstract) | `a [MASK] é uma espécie de` $X_1$ | 0.24 | 0.56 | 0.08 | 0.56 |
| Hyponym-of (abstract) | $X_1$ `é um tipo de [MASK]` | 0.14 | 0.52 | 0.08 | 0.38 |
| | $X_1$ `ou outro [MASK]` | 0.14 | 0.42 | 0.08 | 0.38 |
| Hyponym-of (concrete) | $X_1$ `é um tipo de [MASK]` | 0.54 | 0.88 | 0.28 | 0.56 |
| Part-of | `um [MASK] tem` $X_1$ | 0.12 | 0.30 | 0.06 | 0.28 |
| Has-Part | `um [MASK] é uma parte de` $X_1$ | 0.12 | 0.28 | 0.06 | 0.24 |

**Table 1.** Patterns that outperform LRCos.

Results suggest that, even though a pre-trained BERT is not ready for being directly used in the automatic enrichment of LKBs, it is a great source of such knowledge. Transforming it to explicit relation instances, e.g., represented in RDF, does not require fine-tuning and is mostly a matter of finding the right lexical patterns. Moreover, for better accuracy, results may be further filtered by humans or by dedicated automatic procedures.

As in previous work for English [3], we confirmed that BERT works particularly well for the acquisition of hypernyms, especially if concrete concepts are involved[2]. On the other hand, no tested pattern outperformed LRCos for hypernym between verbs, nor for synonymy between nouns, verbs or adjectives. The main reason for this is the lack of patterns identified for these relations. Moreover, as other studies have shown [7], for synonymy, LRCos leads to minimal to no improvements, when compared to simply computing the cosine similarity.

Future plans include: (i) improving accuracy by combining several patterns (e.g., including longer patterns, acquired from corpora [1]) and ranking measures; (ii) analysing how well sentence probability correlates with relations prototypicality, e.g., approximated by the number of resources where a relation instance is found [6].

## Bibliography

[1] Bouraoui, Z., Camacho-Collados, J., Schockaert, S.: Inducing relational knowledge from BERT. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7456–7463. AAAI Press (2020)

[1] Predictions for one entry had to maximise the similarity with the word in the first column × the probability of belonging to the class of words in the second column, given by a Logistic Regression classifier trained in all other entries.

[2] In TALES, entries for hypernymy and hyponymy are split between those involving more concrete and more abstract concepts.

[2] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4171–4186. ACL (2019)

[3] Ettinger, A.: What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the ACL 8, 34–48 (2020)

[4] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)

[5] Gladkova, A., Drozd, A., Matsuoka, S.: Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In: Procs of NAACL 2016 Student Research Workshop. pp. 8–15. ACL (2016)

[6] Gonçalo Oliveira, H.: A survey on Portuguese lexical knowledge bases: Contents, comparison and combination. Information 9(2) (2018)

[7] Gonçalo Oliveira, H., Sousa, T., Alves, A.: TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In: Procs of ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020). CEUR Workshop Proceedings, vol. 2693, pp. 41–47. CEUR-WS.org (2020)

[8] Hartmann, N.S., Fonseca, E.R., Shulby, C.D., Treviso, M.V., Rodrigues, J.S., Aluísio, S.M.: Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In: Proc of 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017) (2017)

[9] Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of 14th Conference on Computational Linguistics. pp. 539–545. COLING 92, Association for Computational Linguistics, Morristown, NJ, USA (1992)

[10] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proc of Workshop track of ICLR (2013)

[11] Paes, G.E.: Detecção de Hiperônimos com BERT e Padrões de Hearst. Master's thesis, Universidade Federal de Mato Grosso do Sul (2021)

[12] Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Procs of 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543. EMNLP 2014, ACL (2014)

[13] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.: Language models as knowledge bases? In: Proc 2019 Conf on Empirical Methods in Natural Language Processing and 9th Intl Joint Conf on Natural Language Processing (EMNLP-IJCNLP). pp. 2463–2473. ACL (2019)

[14] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog 1(8), 9 (2019)

[15] Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: Proc of Brazilian Conf on Intelligent Systems (BRACIS 2020). LNCS, vol. 12319, pp. 403–417. Springer (2020)

# Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus

Vasile Păiș[1[0000−0002−0019−7574]] and Maria Mitrofan[1[0000−0001−7466−2013]]

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania {vasile,maria}@racai.ro
https://www.racai.ro

## Extended Abstract

In the context of the recent international project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL)[1] a large comparable corpus of legal documents for 7 languages (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, Slovenian) was created [7]. This includes a monolingual sub-corpus for the Romanian language [6]. The Romanian corpus, as well as the other MARCELL corpora, was split at sentence and token level, lemmatized, and annotated at token level. Annotations comprise part-of-speech tags, dependency parsing, named entities and finally the corpus was enriched with IATE and EUROVOC terminologies. Named entities were identified using a general-purpose tool [3], available at that time for the Romanian language. The tool was not trained on any legal texts.

Previous studies [1] have shown that named entity recognition (NER) plays an important role in machine translation. Initial evaluation of the Romanian NER system on the MARCELL sub-corpus (as reported in [6]) provided rather modest results (an overall precision of 64.1%). This made us consider improving the recognition performance by (a) constructing a manually annotated corpus in the legal domain and (b) a fine-tuned domain-specific NER system. This work presents an overview of the created gold corpus and initial experiments in creating a NER system for the Romanian legal domain.

The LegalNERo [4] corpus was constructed with the help of 5 annotators under the supervision of two researchers with experience in corpus annotation. The entities considered are: persons, locations, organization, time and legal references. There are 17,429 total entities, grouped in 370 documents, comprising 8,284 sentences. Inter-annotator agreement provided good results, with an average Coehn's Kappa of 0.89. The released corpus contains annotations at text-span and token levels. Locations were marked with GeoNames codes, where an automatic identification was possible. The resulting corpus was assembled in RDF format, specific to linguistic linked open data, including all the available annotation levels. The corpus is freely available for download[2] and it can be accessed online using a Sparql endpoint[3]. The Sparql endpoint

[1] https://marcell-project.eu/
[2] https://doi.org/10.5281/zenodo.4772094
[3] https://relate.racai.ro/datasets/dataset.html?tab=query&ds=/legalnero

allows easy extraction of entities from the corpus, useful for creating gazetteer resources for NER systems.

Initial experiments using NeuroNER [2] produced two models: (a) for recognizing all the entities, yielding an F1 score of 84.00 %, and (b) for recognizing only persons, locations, organizations and time expressions in the legal domain, yielding and F1 score of 84.70%. For constructing the models we used word embeddings [5] constructed using the Representative Corpus of Contemporary Romanian Language (CoRoLa). These models form a baseline for further fine-tuning and creating improved NER models for the Romanian legal domain. Additional experiments, with different neural architectures as well as different embedding representations, are currently under way. Until better models will become available, the current baseline models are available for online usage from the RELATE platform [4].

## Bibliography


[1] Babych, B., Hartley, A.: Improving machine translation quality with automatic named entity recognition. In: Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003 (2003)

[2] Dernoncourt, F., Lee, J.Y., Szolovits, P.: NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. Conference on Empirical Methods on Natural Language Processing (EMNLP) (2017)

[3] Păiș, V.: Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language. Ph.D. thesis, Romanian Academy (2019)

[4] Păiș, V., Mitrofan, M., Gasan, C.L., Ianov, A., Ghiță, C., Coneschi, V.S., Onuț, A.: Romanian Named Entity Recognition in the Legal domain (LegalNERo) (May 2021). https://doi.org/10.5281/zenodo.4772094

[5] Păiș, V., Tufiș, D.: Computing distributed representations of words using the CoRoLa corpus. Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science **19**(2), 185–191 (2018)

[6] Tufiș, D., Mitrofan, M., Păiș, V., Ion, R., Coman, A.: Collection and annotation of the Romanian legal corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 2773–2777. European Language Resources Association, Marseille, France (May 2020)

[7] Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J.: The MARCELL legislative corpus. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 3761–3768. European Language Resources Association, Marseille, France (May 2020)


[4] `https://relate.racai.ro/index.php?path=ner/demo`

# Investigating Academic Document Structure using Object Detection Methods*

Margaux Susman[1], Djuddah Leijen[2],
Nicholas Groom[3], and Christer Johansson[1]

[1] University of Bergen, Bergen, Norway
[2] University of Tartu, Tartu, Estonia
[3] University of Birmingham, Birmingham, United Kingdom
Margaux.Susman@uib.no
djuddah.leijen@ut.ee
n.w.groom@bham.ac.uk
Christer.Johansson@uib.no

**Abstract.** The rhetorical structure of academic research papers written in English is now well understood, much less is known about the generic conventions governing academic texts written and published in less-studied languages. This article investigates the automatic detection of rhetorical patterns in academic texts using machine learning algorithms which were originally designed for image object detection purposes, and are thus entirely language independent. Our initial results indicate that this graphical, image-based approach to genre analysis is feasible. We intend to extend our approach to the detection of local variants and the rules of those variants.

**Keywords:** Object detection · Deep learning · Academic writing.

## 1 Introduction

Our understanding of rhetorical structures in academic texts took a giant leap with the rise of the genre-based approach pioneered by Swales [1]. This understanding, however, is substantially limited to academic writing in English; much less is known about traditional or emerging patterns in other languages. The Bwrite project aims to address this gap by investigating the structural properties of academic texts (BA, MA, PhD theses, and published scientific work) published in Estonian, Latvian, and Lithuanian. By studying large numbers of texts, we aim to detect patterns on three levels: macro- (whole text), meso (e.g. paragraphs) , and micro-level (sentence level).

In this paper, we focus on detecting the structure of documents at the macro level. We investigate whether the internationally standardized IMRaD (Introduction, Methodology, Results and Discussion) format is also prevalent in academic papers published in these three Baltic countries, or whether other structures emerge.

* The Bwrite project (EMP475) is funded by Iceland, Liechtenstein and Norway through the EEA Grants and Norway Grants.

## 2 Document structure: looking, not reading

### 2.1 Dataset and challenges

Our database was built from publicly available texts from various universities and consists of full texts in PDF format. This format gives us two possibilities, namely, transforming the PDFs to text or changing the documents to images. The first option ignores information (e.g. font, font size, layout, etc.) that may be of fundamental value for the detection of document structure. Consequently, we selected the second option and considered the documents as images. Applying object detection methods means that the language in which the text is written is not relevant; as one does not need to understand the text to determine the structure of the document, the algorithm focuses exclusively on the aesthetics of the page layout. Accordingly, the algorithm can be used on any language, be it English, Estonian or Tamil. We thus transformed our documents into sets of images, in which one page is equal to one image.

### 2.2 Pre-processing and methods

We used Open Labeling [2] to manually annotate the training and validation datasets. The training set contained a thousand images. The validation set was made of 318 images. These images were pages of academic writing papers (mostly theses), originating from diverse fields of study and publicly available online. We drew bounding boxes around areas of interest, namely headers, tables of contents, titles and body (the latter consisted of paragraphs, tables, figures, etc.). We then applied the YOLO algorithm developed by Redmon et al. [3, 4] to analyse the document layouts.

YOLO is a deep learning model, which is trained to draw bounding boxes around regions of interest. Concomitantly, YOLO estimates the probabilities of a specific category to be combined with a bounding box. YOLO performs these tasks using a convolutional neural network. The algorithm also authorizes multi-label classification permitting the overlap of many different categories (e.g. we have a ”body” category which contains everything that is not a section header, a table of content or a title. If one is interested in tables and figures, the ”body” category would cover the ”tables” and ”figures” categories but YOLO would still be able to understand that a same object (e.g. a table) could also be part of a larger object (e.g. the body)).

## 3 Primary results and a brief glimpse at the next steps

The algorithm obtained promising results on the validation set, with a mean Average Precision of 0.990 [5] on the correct prediction of the image windows. Once YOLO has made its predictions on our data, we used the coordinates from the algorithm to extract the regions of interest on each page. Images are heavier than text: as we treat thousands of documents, we make use of an autoencoder to reduce their weights. Unlike typical neural networks which find a function mapping $x$, a feature, to $y$, its category, autoencoders find the function that maps a feature $x$ to itself. An image is transformed into a vector of numerical values in the encoding stage and, in the decoding stage, the algorithm takes as input a vector from the encoder and returns an image as close to the original document as possible. Moreover these vectors give importance to the position of the headers in the document, we use them to pursue with the classification.

## Bibliography

[1] Swales, J. (2014). Create a research space (CARS) model of research introductions. Writing about writing: A college reader, 12-15.
[2] Cartucho, J., Ventura, R., Veloso, M.: Robust object recognition through symbiotic deep learning in mobile robots. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2336–2341, Madrid, Spain (2018)
[3] Redmon, J., Divvala, R., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. Las Vegas, US (2016)
[4] Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. CoRR abs1804.02767, (2018)
[5] Measuring Object Detection models - mAP - What is mean Average Precision?, https://towardsdatascience.com/what-is-map-understanding-the-statistic-of-choice-for-comparing-object-detection-models-1ea4f67a9dbd. Last accessed 29 Jun 2021