# Czech Science Foundation - Part C

**Applicants:** Ing. Marko Sahan, Ing. Vladislav Belov

**Project title:** Coreference Resolution for Improved Natural Language Understanding

## 1 Motivation

Modern Natural Language Processing (NLP) approaches are able to achieve significant results in standard textual analysis tasks. The list of tasks includes but is not limited to such fields as text classification, e.g. determining the general topic of the news article [1] or determining text author's attitude towards the topic [14]; sequence tagging, e.g. named entity recognition (NER) [19, 28, 25, 11] and part-of-speech tagging [2]; and text generation [6]. For some applications it is important to combine these tasks to achieve more comprehensible results. For instance, in the general case of the sentiment analysis task one aims to classify whether the author of the piece of text refers to the topic in a negative or positive sense. However, to obtain a finer understanding of why their attitude is inferred to be some particular value, it is important to discern contextual dependencies within the piece, especially in cases when the range of output values goes beyond "polarity" (positive/neutral/negative) and matures into a broader spectrum of values like doubt, contempt, or enjoyment. The NLP research community not only actively develops models for improved natural language understanding, i.e. representing language in a vector space, [16, 5, 26] but also proposes different fine-tuning approaches for these models [10, 7, 17, 3].

In this project we aim to perform research in construction of better textual dependencies for the task in the form of improved Coreference Resolution (CR). The CR task is very complex to solve and from 2019 to this date [**?**] no improvement was achieved on the standard benchmark [23]. The recent article [20] achieved the same metrics as the one from 2019. Coreference resolution combines detection and linking of various mentions of entities within the text: linking noun phrases with their counterparts and pronouns, anaphora disambiguation, linking words with their pro-forms, etc. These models have a significant impact on the quality of the text mining algorithms. A good use case where coreference resolution can be applied is categorization of entities and their pronouns in order to provide one with a wider spectrum of information for future decision making. Based on the extracted data it is possible to unify all knowledge in a form of a Knowledge Graph (KG) [22]. The dependencies and connections between the entities can be used for enriching the feature space with the highly discriminative samples for the further tasks. For example, let us assume that we have two consecutive sentences like "John Smith and Amanda Brown are accountants in XYZ company. Amanda's colleague was accused of drunk driving". Based on these sentences we would like to classify if some of the entities from the text can be charged for a misdemeanor. For a human reader it is obvious that Amandas's colleague refers to John. However for a machine that is a very hard task. However, proper identification of entity clusters like John Smith, Amanda's colleague, Amanda Brown, XYZ would greatly improve the machine's understanding of the piece of text.

In the scope of this work, we expect to improve the current state of the art (see the following section) by means of its further augmentation. Firstly, we believe that advancement can be achieved via the modification of the existing CR-solving model which is applied on top of vector embeddings. Since the model relies on scoring entity mentions and clustering them, significant changes can be brought with nonlinear dimensionality reduction which, in neural-network-based structures, can be achieved by means of autoencoders [27, 18]. In addition to that, initial tests show that advancements from the named entity recognition field may also provide us with meaningful results, as NER models also focus on entities and context surrounding them: in this case we propose to explore conditional random fields [19, 28] as one of potential candidates. In addition to that, we wish to entertain the possibility of integration of uncertainty [], as it has been shown that this approach may enhance the learning process significantly.

As another output of this project we expect to not only push the boundaries of the state-of-the-art algorithms but also introduce the first Czech and Slovak Coreference Resolution dataset that can be used as a new benchmark for the Czech and Slovak CR models evaluation. Based on the new data and better Coreference Resolution approach we would like to generalize the CR algorithm as a Multilingual solution.

## 2 State of the Art

Modern Coreference Resolution (CR) algorithms are combinations of sophisticated vector embeddings representing context and deep neural network superstructures that perform the coreference resolution itself.

Natural language understanding (NLU) models. The set of existing models for NLU is vast. Arguably, one of the most prominent points in history of such models is when the continuous bag-of-words and skip-gram approaches were introduced [15]. At that point machines started to be able to learn the context surrounding particular words and their vector representations acquired the ability to represent this context, meaning proximity of such vectors in terms of a metric of choice (L2, cosine/angular similarity) veritably described similarity of words or contexts. Still, models of these types were far from perfect, as they provided one with constant vectors per word for a pre-set vocabulary. Context-dependent representations with flexible vocabularies became available thanks to the introduction of the Transformer architecture [21] applied on the vocabulary formed not only by words but also by character n-grams constructed as meaningful parts of words. The power of the Transformer architecture lies in its encoding and decoding capability improved by the self-attention mechanism which learns to put stress on parts of text sequences. This gave birth to a lot of transformer-based language models such as the Bidirectional Encoder Representations from Transformers (BERT) [5], its fine-tuned variations [8, 10] and further models [16, 4]. To this date, SpanBERT [7] has proven to be the most efficient architecture for coreference resolution. Its crucial difference from the standard BERT model is that it learns to predict the content of masked spans of text, taking into account their beginnings and endings, omitting the ability of the base BERT model to predict foregoing sentences, whereas BERT learns to predict the following sentence for each preceding one and attempts to infer individual masked tokens.

The first end-to-end coreference resolution model was introduced in [9]. Its crucial difference from its predecessors was that it did not require preprocessing in the form of syntactic parsing or rule-based mention detection, since the model is able to learn mention dependencies on its own to a forerunner-outperforming extent. The main idea of the model is to learn to score pairs of textual spans in such a way that takes into account, firstly, if these spans are entity mentions and, secondly, whether the pair is of type antecedent-descendant in terms of coreference. Span representations are provided by the NLU model of choice. The goal is to be able to assign to each span an antecedent span. The current state-of-the-art approach [7] utilizes the same structure on top of SpanBERT. One of the crucial drawbacks of the scoring approach is the choice of spans: sizes of relevant spans can be different so a constant width of the window may not always be the right choice; spans can either overlap or be disjoint; if they overlap, the value of how large the overlap is also becomes a hyperparameter. In addition to that, the number of scoring procedures is quadratic in complexity: each span has to be scored against every its counterpart. If the length of the document is large, the memory needed to store all entity mentions may become an issue (in [24] authors propose an incremental structure for the CR model which needs a lot less memory for the price of a slight decrease in performance).

The CR scoring model takes as input sequences of high-dimensional word-vectors which were produced by multiple consecutive nonlinear mappings. For that reason, one can assume the resulting language-representing structure is highly nonlinear and noisy. In such cases, nonlinear dimensionality reduction helps to preserve crucial information, e.g. probability distribution [12] or structure of neighborhood-based local metric spaces [13], while reducing the dimension of the space significantly. Learningimensionality reduction We aim to improve the model by means of

# References


[1] B. Altinel and M. C. Ganiz. Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, 54(6):1129–1153, 2018.

[2] B. Bohnet, R. McDonald, G. Simões, D. Andor, E. Pitler, and J. Maynez. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder, 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang. Long text generation via adversarial training with leaked information, 2017.

[7] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. *CoRR*, abs/1907.10529, 2019.

[8] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[9] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045, 2017.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[11] J. Luoma and S. Pyysalo. Exploring cross-sentence contexts for named entity recognition with bert, 2020.

[12] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[13] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[14] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[16] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018.

[17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[18] R. Sahay, R. Mahfuz, and A. E. Gamal. Combatting adversarial attacks through denoising and dimensionality reduction: A cascaded autoencoder approach. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, 2019.

[19] J. Straková, M. Straka, and J. Hajič. Neural architectures for nested ner through linearization, 2019.

[20] S. Toshniwal, S. Wiseman, A. Ettinger, K. Livescu, and K. Gimpel. Learning to ignore: Long document coreference with bounded memory neural networks, 2020.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[22] Q. Wang, Z. Mao, B. Wang, and L. Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

[23] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, and et al. Ontonotes release 5.0, 2013.

[24] P. Xia, J. Sedoc, and B. Van Durme. Incremental neural coreference resolution in constant memory, 2020.

[25] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention, 2020.

[26] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[27] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, and S. Marshall. Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing*, 185:1–10, 2016.

[28] J. Zhanming and L. Wei. Dependency-guided lstm-crf for named entity recognition, 2019.