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 [20]; sequence tagging, e.g. named entity recognition (NER) [27, 37, 34, 17] and part-of-speech tagging [5]; and text generation [10]. 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, [23, 8, 35] but also proposes different fine-tuning approaches for these models [15, 11, 24, 6].

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 [11] no improvement was achieved on the standard benchmark [31]. The recent article [28] achieved the same metrics as the one from 2019. The CR task is a top-level task for contextual understanding. Thus, the output F1 score of the state-of-the-art approaches show 79.6%, which is not good enough for the real world application tasks. Hence, the range of possibilities for new solutions is far enough from both exhausting and reaching the saturation point of the research.

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) [30] which can be further utilized for linking concepts represented by textual spans. The dependencies and connections between the entities can be used for enriching the feature space with the highly discriminative samples for the further tasks. 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. Another potential application of coreference resolution lies within the problem of opinion mining in media resources where people frequently freely express their views and opinions. For example, heated discussions may emerge under political news articles. In these discussions, participants refer to subjects of the particular article with, for instance, pronouns. Therefore, proper CR may provide better traction of the audience attitude towards entities from the article by linking comment mentions to them.

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 [36, 25], since they may enable the model to extract meaningful de-noised relationships from the high-dimensional structure. 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 (CRF) [27, 37, 37] and attention [34] as potential candidates, as CRF is capable of improving relationship-decoding capabilities of the model and attention learns to put stress on important parts of textual sequences. In addition to that, we wish to entertain the possibility of integration of uncertainty representations algorithms [12, 9, 32]. The model uncertainty measurement and its representation is done through the empirical estimate and sampling from the model weights distribution. The uncertainty representation approach provides the model with the expanded vision of both model learning and inference. The described technique has shown that such algorithms may enhance the learning process significantly [22].

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 solutiwon.

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 [21]. 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 [29] 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) [8], its fine-tuned variations [13, 15] and further models [23, 7]. To this date, SpanBERT [11] 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 [14]. 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 [11] 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 [33] 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 [18] or structure of neighborhood-based local metric spaces [19], while reducing the dimension and noisiness of the space. Therefore, by learning vector representations with more informative dimensions with, for example, such neural structures as autoencoders [36, 25] can lead to improved cluster mention classification results. In addition to that, the sequential nature of text and the need to classify relations between entity mentions puts CR close to the field of NER where, for instance, CRF [27, 37] and attention [34, 4] have proven to be good additions to the sequence-labeling model in the endeavor to learn correct token labels, which is a problem with combinatorial complexity: CRF is capable of decoding of sequential relationships into meaningful labels, which in case of CR may represent relationships between mention spans, whereas attention is able to lay emphasis on trigger words or spans that are essential for the task of question.

Another way to improve the model performance from a different aspect is the addition of the uncertainty measurement. The grand research from Google [22] shows that the point wise estimate of model's parameters does not usually result in an optimal approach. The models uncertainty measurement through the estimate of the empirical model weights distribution has already shown great results in the active learning fields both in Computer Vision [9] and NLP in such tasks like NER [26, 16], text classification [2] and other applications. Thus, the model parameters estimated based on uncertainty algorithms: i) deep ensembles [12] consisting of N networks trained in parallel from different initial conditions; ii) MC Dropout [9] which is an extension of the ordinary dropout that samples binary mask multiplying output of a layer, hence stopping propagation through all neurons where zeros is sampled through the network, the extension applies the sampled mask even for predictions generating samples from the predictive distribution; iii) Stochastic Gradient with Langevin Dynamics (SGLD) [32] which adds additional noise to the gradient in stochastic gradient descent, will bring more efficient training and predictions. Back in 2017, the deep ensembles approach for a CR problem [14] showed a significant increase (3% F1) in the output metrics which confirms the veracity of the concept.

We aim to utilize the knowledge of these aforementioned fields and adapt it to the CR problem to improve the existing framework.

3 Work Packages

WP1: State-of-the-Art Models Optimization and Fine-tuning

T1.1 Further enhancement of the state-of-the-art models in PyTorch (*Marko Sahan*) and TensorFlow 2 (*Vladislav Belov*) to support cluster computation optimization and generalization for various text representation mechanisms.

T1.2 Creation of an environment which is capable of assessing and analyzing the quality of the resulting models in a meaningful and comparable way. (*Vladislav Belov, Marko Sahan*)

WP2: Active-Learning-powered Multilingual Data Corpora Creation

The datasets used for coreference resolution are very limited. The most well known dataset that is used for coreference resolution is OntoNotes 5 [31] and it is only available in English, Chinese and Arabic languages. Thus, another objective of the research is the creation of Czech and Slovak datasets.

We would like to introduce the Active Learning data corpuses creation. The experiment is based on the "smart" way of the data collection when the uncertainty based algorithm will choose the unlabeled data by itself and transfer the instances to the annotators for obtaining the labels.

T2.1 Prepare the active learning data collection and labeling environment. Modern approach of the data collection is based on the random selection of the documents that may meet some criteria like duplicates removal, etc.. The data for the Czech and Slovak corpus are going to be extracted from online news media and czech wiki servers.

Recent approaches showed that it is possible to select the data iteratively with the help of the uncertainty based model feedback [9, 16]. The following approach brings the higher quality data selection that also results in less labeling iterations and lower financial expenses. (*Marko Sahan, Vladislav Belov*)

T2.2 Creation of Czech and Slovak Data Corpuses. The annotations are going to run in the prepared environment from the previous task. In order to reduce the amount of human efforts we will bootstrap the data for the annotations. The bootstrapping approach is based on pre-labeling the Czech and Slovak data with the usage of existing superstructure CR model that uses language independent embeddings for Czech and Slovak language encodings. Described approach will let the annotators to get on average, only task relevant data. In the first iteration, the target estimate is to have at least 30k-50k annotated sentences for the Czech language. In the second iteration Slovak language data corpus will be collected. The target estimate regarding the annotations is the same as for Czech language (30k-50k). Marko Sahan has the experience of partially leading the end to end 20k sentences corporate corpus creation with entity-to-entity information link labels. (*Marko Sahan, Vladislav Belov*)

T2.3 Creation of Multilingual validation Data Corpuses. Modern algorithms have a clear traction towards the language independent approaches like LASER [3]. Thus, the generalization of a coreference task is obvious. Language independent embeddings approach is a representation of the same semantic structures with the same, or at least similar (with respect to a specific metric) vectors. Hence, models trained in English must be able to work with different languages. However, in practice, if the model is not fine tuned with another language that it makes predictions for, the error rate of the predicted instances will be higher (this is the reason for T2.1). Nevertheless, modern state of the art approaches were not tested or evaluated on the multilingual datasets either due to the lack of the quality data or due to the fact that the CR field is relatively young. We expect to create 2k testing data corpuses for French, German, Spanish and Turkish. (*Marko Sahan, Vladislav Belov*)

WP3: Multilingual Coreference Resolution Model

WP3 T3.3 and T3.4 are strongly dependent on WP1 and WP2. However WP3 T3.1 can be done in parallel with the WP2 annotation procedure.

T3.1 Investigation of influence of (nonlinear) dimensionality reduction on mention clusters in an attempt to reduce the number of noisy dimensions and improve shapes of these clusters. We will experiment with autoencoders [36, 25] to achieve a nonlinear mapping of the language-representing manifold onto a lower-dimensional denoised space within a neural-network model. Moreover, since the models for named entity recognition (NER) also work with entity mentions, examination of superstructures designed for NER, e.g. Conditional Random Fields (CRF) [27, 37] and entity-aware attention [34], applied on the coreference resolution problem will be further explored. (*Vladislav Belov*)

T3.2 The grand research from Google [22] shows that the point wise estimate of model's parameters does not usually result in an optimal approach. Based on our preliminary results, we believe that the models' parameter distribution estimate based on uncertainty algorithms e.g. deep ensembles [12], MC Dropout [9] or SGLD [32] will bring more efficient results. Back in 2017, the deep ensembles approach for a CR problem [14] showed a significant increase (3% F1) in the output metrics. Thus, we expect to bring more significant improvements with addition of more sophisticated uncertainty based algorithms. (*Marko Sahan*)

T3.3 Generalization to multilingual CR superstructure via usage of multilingual NLU models will allow us to evaluate (retrain and test) the CR superstructure from WP1 on the data from WP2 both for base state-of-the-art models and improved models with potentially a more efficient architecture and the uncertainty representations algorithms. The environment prepared in WP1 will enable us to retrain and to get the results consistently and in faster timelines. (*Vladislav Belov, Marko Sahan*)

T3.4 The data from T3.3 will allow us to understand the metric baseline for the Multilingual CR model. In the first step, we expect the results from this step to be sufficient enough for their publication as so far no Deep Learning CR results are available for Czech and Slovak languages. (*Marko Sahan, Vladislav Belov*)

4 Research Team

Ing. Marko Sahan

Ing. Vladislav Belov

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] Bang An, Wenjun Wu, and Huimin Han. Deep active learning for text classification. In *Proceedings of the 2nd International Conference on Vision, Image and Signal Processing*, pages 1–6, 2018.

[3] Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[5] 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.

[6] 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.

[7] 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.

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

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

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

[11] 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.

[12] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[13] 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.

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

[15] 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.

[16] David Lowell, Zachary C Lipton, and Byron C Wallace. Practical obstacles to deploying active learning. *arXiv preprint arXiv:1807.04801*, 2018.

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

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

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

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

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

[22] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

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

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

[25] 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.

[26] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

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

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

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

[30] 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.

[31] 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.

[32] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

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

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

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

[36] 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.

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