Czech Science Foundation - Part C

Applicant: Ing. Radek Mařík, CSc.

Project title: Coreference Resolution for Improved Natural Language Understanding

1 Motivation

Modern Natural Language Processing (NLP) approaches can 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 [AG18] or determining text author's attitude towards the topic [MHK14]; sequence tagging, e.g. named entity recognition (NER) [SSH19, ZW19, YAS+20, LP20] and part-of-speech tagging [BMS+18]; and text generation [GLC+17]. 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 an adequate understanding of why their attitude is inferred to be some particular value, it is essential 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, [RN18, DCLT19, YDY+20] but also proposes different fine-tuning approaches for these models [LOG+19, JCL+19, RWC+19, BMR+20].

In this project, we aim to perform research on the construction of better textual dependencies for the task in the form of improved Coreference Resolution (CR). The CR task is very complex to solve. From 2019 to this date [JCL+19] no improvement was achieved on the standard benchmark [WPM+13]. The recent article [TWE+20] achieved the same metrics as the one from 2019. The CR task is a top-level problem for contextual understanding in terms of complexity. Nonetheless, the output F1 scores of the state-of-the-art approaches reach 80.3%, which is not sufficient for real-world applications. Hence, the range of possibilities for new solutions is far enough from both exhausting and landing the saturation point of the research.

Coreference resolution combines detection and linking various mentions of entities within the text: linking noun phrases with their counterparts and pronouns, anaphora disambiguation, linking words with their pro-forms, etc. Hence, CR-solving models significantly impact the quality of the text mining algorithms. A good use case where coreference resolution can be applied is categorizing entities and their pronouns to provide one with a broader spectrum of information for future decision making. Based on the extracted data, it is possible to unify all knowledge in the form of a Knowledge Graph (KG) [WMWG17] which can be further utilized for linking concepts represented by textual spans. Dependencies and connections between the entities can enrich the feature space with highly discriminative samples for other tasks. For example, assume that we have the following two consecutive sentences: "John Smith and Amanda Brown are accountants in XYZ company. Amanda's colleague was accused of drunk driving". Based on these sentences, one would wish to classify if some of the entities from the text can be charged for a misdemeanor. For a human reader, it is evident that Amanda's colleague refers to John. However, for a machine, that is a challenging task. Therefore, proper identification of entity clusters like John Smith, Amanda's colleague, Amanda Brown, XYZ would significantly 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's 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 utilizing its further augmentation. Firstly, we believe that advancement can be achieved via modifying 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 *utilizing autoencoders* [ZRZ+16, SMG19], 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)* [SSH19, ZW19, ZW19] *and attention* [YAS+20] 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. As a further branch of research, we will integrate enhanced models uncertainty measurement algorithms e.g deep ensembles [LPB16], MC Dropout [GIG17], SGLD [WT11] or Vadam [KNT+18] to the coreference resolution superstructure for *the empirical model weights distribution estimate*. The additional information, given the distribution of predicted labels, allows faster model learning (hot and warm start methods) with a lower number of training data, given the model prediction uncertainty [SSM21]. The architecture agnostic generalization of the empirical model weights distribution estimate will grant us more freedom in choosing the CR task superstructure model with preservation of the precise insight into the model processes given prediction-based uncertainty. The model uncertainty measurement and representation are done through the empirical estimate and sampling from the model weights distribution. The uncertainty representation approach provides the model with an expanded vision of both model learning and inference. The described technique has shown that such algorithms may enhance the learning process significantly [OFR+19].

As another output of this project, we expect to not only push the boundaries of the state-of-the-art algorithms but also create a new Czech and introduce the first Slovak Coreference Resolution datasets generated in an active-learning-powered environment. We want to generalize the CR algorithm as a multilingual solution based on the new data and a better Coreference Resolution approach.

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.

The set of existing models for natural language understanding (NLU) is vast. Arguably, one of the most prominent points in the history of such models is when the continuous bag-of-words and skip-gram approaches were introduced [MCCD13]. At that point, machines started to learn the context surrounding particular words. 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 [VSP+17] applied to 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) [DCLT19], its fine-tuned variations [LCG+20, LOG+19] and further models [RN18, CYK+18]. To this date, SpanBERT [JCL+19] has proven to be one of the most efficient architectures 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 preceding sentences. In contrast, BERT learns to predict the following sentence for each preceding one and

attempts to infer individual masked tokens. Another model worthy of mention is a Longformer, whose architecture is based on transformers capable of processing long documents up to 4096 tokens [BPC20].

The first end-to-end coreference resolution model was introduced in [LHLZ17]. 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 can 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. The NLU model of choice provides span representations. The goal is to assign to each span an antecedent span. [JCL+19] belongs to the state-of-the-art approaches which utilize 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 the overlap 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 [XSD20] authors propose an incremental structure for the CR model, which needs a lot less memory for the price of a slight decrease in performance). While previous models are able to achieve decent results, their memory footprint is significant. The authors of [KRL21] bypassed the need to create span representations, relying on a combination of bilinear functions applied on endpoint token representations. In addition to that, the new model is built on top of a Longformer encoder capable of processing long documents.

The CR scoring model takes as input sequences of high-dimensional word-vectors 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 [Maa14] or structure of neighborhood-based local metric spaces [MHM18], while reducing the dimension and noisiness of the space. Therefore, learning vector representations with more informative dimensions with, for example, such neural structures as autoencoders [ZRZ+16, SMG19] can lead to improved cluster mention classification results. In addition to that, the sequential nature of the text and the need to classify relations between entity mentions puts CR close to the field of NER where, for instance, CRF [SSH19, ZW19] and attention [YAS+20, BCB14] 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 can emphasize trigger words or spans that are essential for the task of question.

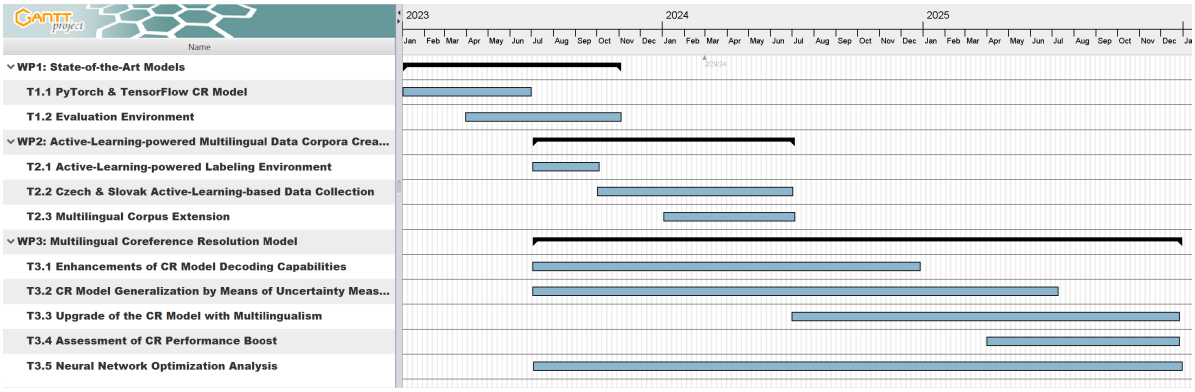
Another way to improve the model performance from a different aspect is the addition of the uncertainty measurement. The grand research from Google [OFR+19] shows that the point-wise estimate of model parameters does not usually result in an optimal approach. The models uncertainty measurement through the estimate of the empirical model weights distribution has already shown remarkable results in the active learning fields both in Computer Vision [GIG17], and NLP in such tasks like NER [SYL+17, LLW18], text classification [AWH18] and other applications. Thus, the model parameters are estimated based on uncertainty algorithms: i) deep ensembles [LPB16] consisting of N networks trained in parallel from different initial conditions; ii) MC Dropout [GIG17] which is an extension of the ordinary dropout that samples binary mask multiplying the 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) [WT11] 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 [LHLZ17] 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 fields as mentioned above and adapt it to the CR problem

to improve the existing framework.

For languages other than English, the state of art for the CR is arguably even farther behind. For example, coreference resolution for Czech was attempted on the PCEDT dataset [HHP+12] in [Nov17]. However, the overall performance of the approach did not reach the mark of 0.7 in terms of F-score. In addition to that, no transformer-based NLU model was available at the time. We aim to address this in our research, as SpanBERT and Longformers demonstrate superior performance to other neural supervised techniques.

3 Work Packages



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

T1.1 Enhanced State-of-the-Art CR Models (Belov, Sahan) Further enhancement of the state-of-the-art models in PyTorch to support cluster computation optimization and generalization for various text representation mechanisms. The existing available implementations are not flexible for experimentation. They are focused on a specific architecture with specific encoding models; some even utilize the Static-Computational-Graph approach. However, modern technologies, e.g., PyTorch and TensorFlow 2, employ Dynamic Computational Graphs [NDG+17].

T1.2 Evaluation Environment (Belov, Sahan) The task of coreference resolution is complex and its performance is evaluated with numerous metrics [VBA+95, BB98, Luo05], as the standardly used F1 score [CL06] does not embrace all aspects of the discriminative power of features. Therefore, creation of an environment which is capable of assessing and analyzing the quality of the resulting models in a meaningful and comparable way is crucial. The initial structure of the platform is already arranged together with a functioning prototype of [KRL21].

WP2: Active-Learning-powered Multilingual Data Corpora Creation

The number of datasets used for benchmarks of coreference-resolution models is quite limited. The most well-known one is OntoNotes 5 [WPM+13], and it is only available in English, Chinese, and Arabic languages. The recently published multilingual harmonized corpus for Czech and ten other languages [NNP+21] extends the limits for new models in terms of multilingualism. However, our research's objective is to create the new Active-Learning-Boosted Czech and Slovak datasets optimized for Machine Learning purposes, especially benchmarks on our neural models. The benefit of this package is not only in the development of the new corpus but also in applying active learning techniques, which will reduce the number of effort annotators put into their work and improve the quality of the resulting data.

We want to introduce the Active Learning data corpora creation. The experiment is based on the "smart" way of 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 Active-Learning-powered Labeling Environment (Belov, Sahan) Environment preparation for data collection and labeling with active learning. The modern approach to 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 will be extracted from online news media and Czech wiki servers. Recent approaches showed that it is possible to select the data batches iteratively with the help of the uncertainty-based model feedback both for smaller size batches [GIG17, LLW18], and significantly big batches [CDG+21]. The following approach brings higher quality data selection, resulting in fewer labeling iterations and lower financial expenses.

T2.2 Czech & Slovak Active-Learning-based Data Collection (Belov, Sahan) Corpora creation for Czech and Slovak languages. The annotations will run in the prepared environment from the previous task. In order to reduce the amount of human effort, we will bootstrap the data for the annotations. The bootstrapping approach is based on pre-labeling the Czech and Slovak data using the existing superstructure CR model that uses language independent embeddings for Czech and Slovak language encodings. The described approach will let the annotators get only task-relevant data on average. The target estimate is to have at least 30k-50k annotated sentences for the Czech language in the first iteration. In the second iteration, the Slovak language data corpus will be collected. The target estimate regarding the annotations is the same as for the 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.

T2.3 Multilingual Corpus Extension (Belov, Sahan) Modern algorithms display evident traction towards the language-independent approaches like LASER [AS19]. Thus, the generalization of a coreference task is obvious. The language-independent embeddings approach represents 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, the modern state of art approaches were not tested or evaluated on the multilingual datasets either due to the lack of quality data or because the neural CR field is relatively young. In addition to CorefUD [NNP+21], we expect to extend our T2.2 data in the same active-learning-enabled environment with 2K-large corpora for French, German, Spanish, and Turkish for validation purposes.

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. WP3 T3.1 and T3.2 are the core work packages of the proposed project.

T3.1 Enhancements of CR Model Decoding Capabilities (Belov) Investigation of the *(nonlinear) dimensionality reduction influence on mention clusters* in an attempt to reduce the number of noisy dimensions and improve the shapes of these clusters. We will experiment with *autoencoders* [ZRZ+16, SMG19] *to achieve a nonlinear mapping of the language-representing manifold onto a lower-dimensional de-noised 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) [SSH19, ZW19] and entity-aware attention [YAS+20], applied on the coreference resolution problem will be further explored.

T3.2 CR Model Generalization by Means of Uncertainty Measurement (Sahan) Integration of the uncertainty representations algorithms e.g deep ensembles [LPB16], MC Dropout [GIG17], SGLD

[WT11] or Vadam [KNT+18] to *the coreference resolution superstructure for the empirical model weights distribution estimate*. The beforehand mentioned integration *will result in more efficient learning and inference procedures*. The primary investigation is concentrated around faster model learning (hot and warm start methods) with a lower number of training data, given the model prediction uncertainty [SSM21]. The diversity of embedding-based superstructure model architectures makes the models' uncertainty prediction slow and inefficient. The further step in the uncertainty estimate research is a generalization of the uncertainty given different models architectures (e.g., Vadam [KNT+18]). The output of the study will allow us to use the architecture agnostic empirical model weights distribution estimate for better noise measurement, faster learning, and label distribution prediction, specifically for different types of CR task embedding superstructure.

T3.3 Upgrade of the CR Model with Multilingualism (Belov, Sahan) Generalization to multilingual CR superstructure via the 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 state-of-the-art base models and improved models with potentially a more efficient architecture and the uncertainty representation algorithms. The environment prepared in WP1 will enable us to retrain and get the results consistently and faster.

T3.4 Assessment of CR Performance Boost (Belov, Sahan) Assessment of influence produced by approaches proposed in T3.1 and T3.2 utilizing a systematic measurement of relevant for CR metrics [VBA+95, BB98, Luo05]. Detailed evaluation of contribution provided by each of the proposed solutions both on the currently existing benchmark for English, Arabic, and Chinese [WPM+13] and on the newly created Czech & Slovak corpora. Moreover, in the scope of the validation process, we aim to test the multilingual capabilities of the CR approach on the data obtained from T.2.3, which were not accessible to the model during training iterations.

T3.5 Neural Network Optimization Analysis (Marik) Approaches based on neural networks deliver excellent results if their architecture is designed adequately for the solved problem. However, if their architecture does not provide solid gradient paths for optimization convergence, it is not easy to discover its deficiencies causing poor performance. To avoid such events in the project, we will create a monitoring tool and suitable visualizations that should help resolve the issues. To tackle a massive amount of optimized neural network parameters, we will utilize sparse representations such as complex networks.

4 Research Team

The research team consists of a senior researcher leading two Ph.D. students. All three have been working in artificial intelligence research for several years, focusing on natural language processing methods in the last three years. Therefore, the design of the objectives of this project is based on the experience gained in the immensely successful publications and bachelor's, project, or diploma theses of both students. Their work has been directly linked to the methods further developed in the proposal of this project. However, the conditions and the need to address coreference resolution also flow from several projects supported by the GAČR and TAČR agencies, which were led or participated in their solution by their supervisor Radek Mařík, i.e., NLP based projects ČTK News (2019-2021), Beey(2020-2021), Covid19 (2020-2022), Signal and noise (2021-2023), CEDMO fact-checking (2021-2024). We exchange experience with abroad teams (e.g., Josef Kittler, CVSSP University of Surrey, UK; Heikki Kalviainen, CVPRL, LUT, Finland)

Applicant Radek Mařík was a co-applicant of the GACR GA16-07210S project successfully solved with an excellent evaluation. He has been the team member of several successfully solved projects supported by MSM (LTT18007) and MV0 (VI20152020008, VG20102015053 in the last five years). He is currently a team member of two TAČR projects (FW01010468, TL04000176), all focused on processing Czech texts in the field of media news. In project TACR TL02000288, the applicant addresses

the issue of clustering of text fragments, clustering metrics, and the possibility of using community detection from complex networks analysis in clustering, which are used as partial steps in the tasks of generating text messages based on structured data, generating summaries, fact-checking, detection and analysis of the significance of events in the news. In all project cases, tasks encountered the issue of resolution co-referencing, the current state of which, either from a theoretical or practical point of view, does not provide a satisfactory performance and prevents obtaining results with much higher resolution, for example, in the form of knowledge graphs.

The team benefits from Radek Mařík's experience gained during his previous 14 years-long industrially AI-oriented research work at Rockwell Automation and CA Technologies, for example resulting in a patent [Mar11]. This project's theory, methods, and properties have been studied and implemented in more than nine diploma thesis projects, four bachelor theses, dozens of peer-reviewed, WOS / Scopus indexed conferences, and IF journal papers published in the last five years. Additional methods and their properties have been studied during several other research projects dealing with different application domains, including, for example, ProtoSpy focused on monitoring, communication structure reconstruction, safety assessment, and synchronous event detection for network traffic [MBKK15, MKP14] and might create additional views on coreference resolution issues.

The software platform for coreference resolution with selected state-of-the-art techniques has been partially implemented. The proposed project will enable us entirely focus on detected coreference resolution issues with the dedicated effort. Also, it will help us to cover the conference financing.

The applicant, Ing. Radek Mařík, CSc., is a specialist in complex network analysis combined with machine learning and data mining, used to develop AI and natural language processing techniques, both theoretical and practical, specialized and applied in non-traditional domains (media news, journalism, telecommunications, optical sensors, Egyptology, ceramic tile industry). The techniques can also be reused in this project based on a shared abstract mathematical and algorithmic foundation. For example, we have already developed methods for keyword, topic, and phrase detection and indexing, consistency maintenance of input data [Mar16a], automated grouping people into families handling uncertainty and logic, automated huge family tree building, layout, and visualization (various possibilities of visualization, that might be reused in this project for deep neural network internal properties analysis, were implemented in [Mar16b, Mar16c, Mar17b, Mar17a, Mar18b, Mar19], automated detection of families with a significant level of nepotism using techniques of social network analysis and data mining [DM15, DMBC17], detection of strategic titles and powerful officials using information theory, that has also been successfully tested in media news topic aspect detection related to this project work [Mar18a], reconstruction of administration development during the Old Kingdom using hidden Markov models [DMBC17, MC17], specialized methods for community detection in hierarchical multipartite networks capable of uncover coreference relations [MZ19, Bel20, Zik20].

Radek Mařík will manage the project. He supervises both Ph.D. students. He will also be responsible for evaluating neural network representation efficiency and their visualizations with the support of complex network analysis methods. Thus, he will deliver feedback on possible improvements to both students.

Ing. Marko Sahan, a Ph.D. candidate, has focused on active learning for text classification with enhanced uncertainty representation in his diploma thesis [Sah20]. The significant contribution of his work has resulted in the best student paper at the ECNLPIR 2021 conference [SSM21]. He will explore mainly coreference resolution improvements using advanced uncertainty representations.

Ing. Vladislav Belov, a Ph.D. candidate, has studied Manifold Learning, Relation Mining, and Kernel Density Estimation [Bel20]. His work resulted in contributions to CIIS 2021 [BM21a] and ACAI2021 [BM21b]. He will focus on the influence of nonlinear dimensionality reduction and structural relations (CRF, entity-aware attention) to coreference resolution.

The team operates with sufficient software library implementations covering both processing and visualizations and hardware and software resources to perform planned simulations, data mining, and advanced neural network-based processing. We have access to the RCI server (48x NVIDIA Tesla V100) at the CTU, which we use for large batch (deep) machine learning and classification. The planned acquisition of new hardware components will make the research even more efficient. The new invested notebook (replacing a similar computer out of guarantee) will enable the analysis of powerful techniques of state-of-art deep learning to detect deficiencies using complex network analysis with sufficient visualization support. Both new notebooks will be used for fast script prototyping; small batch development processing; input data conversions, analysis, and preparation for extensive batch processing by the RCI server; and output data analysis and visualization, and other tasks that are more difficult or not efficient to deploy to the RCI server. The notebooks will also enable mobile result presentations to our clients (e.g., media news houses, journalists) and will make feasible home office arrangements. However, we will also rely on the current hardware and laboratory equipment available at the CTU. Of course, all team members have skills in programming and new rapid method design and implementation, and volume data processing. Diploma thesis students will also implement extensions, improvements, and new algorithms covering required processing.

5 Benefits and Outputs of the Projects

We expect to define new methods for solving coreference resolution with better performance applicable to texts in English and supporting Czech. We believe that the outlined strategy using *uncertainty representation, representation of structural entities, and nonlinear dimensionality reduction will significantly enhance the performance of CR*. The second output will verify the applicability of the proposed techniques in processing media reports and contributions with emphasis on the Czech environment (e.g., we have a ČTK dataset of 4.5 million news, 2.5 million Czech news dealing with COVID-19). Successful deployment would pave the way for a much better, more accurate assessment of the content of media channels and disinformation attempts and open up the possibility of a more objective assessment of their impact on the public based on processing a massive stream of online messages by tracing coreference of entities.

Each year, we expect to deliver a conference paper (conferences dedicated to NLP such as ACL, COLING, NeurIPS, CoNLL, ICML, ICDM). However, the main focus will be given to three impact factor journal papers delivered in the second and third years (e.g., Transactions of the Association for Computational Linguistics, Journal of Computational Linguistics, Journal of Information Retrieval, Journal of Machine Learning).

References

[AG18] B. Altinel and M. C. Ganiz. Semantic text classification: A survey of past and recent advances. *Information Processing & Management*, 54(6):1129–1153, 2018.

[AS19] 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.

[AWH18] 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.

[BB98] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics.

[BCB14] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[Bel20] Vladislav Belov. Nonlinear dimensionality reduction. Master's thesis, CTU in Prague, 2020.

[BM21a] Vladislav Belov and Radek Marik. Manifold Learning Projection Quality Quantitative Evaluation. In *2021 The 4th International Conference on Computational Intelligence and Intelligent Systems*, CIIS 2021, New York, NY, USA, 2021. Association for Computing Machinery.

[BM21b] Vladislav Belov and Radek Marik. Tessellation-Based Kernel Density Estimation. In *2021 4th International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI'21, New York, NY, USA, 2021. Association for Computing Machinery.

[BMR+20] 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.

[BMS+18] 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.

[BPC20] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150, 2020.

[CDG+21] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34, 2021.

[CL06] Yi-Wei Chen and Chih-Jen Lin. Combining svms with various feature selection strategies. In *Feature extraction*, pages 315–324. Springer, 2006.

[CYK+18] 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.

[DCLT19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[DM15] Veronika Dulíková and Radek Mařík. Social network analysis in the Old Kingdom society: the nepotism case. In M. Bárta, F. Coppens, and J. Krejčí, editors, *Abusir and Saqqara 2015, Czech Institute of Egyptology, Charles University, Prague, CZ*, pages 63–83, 2015.

[DMBC17] Veronika Dulíková, Radek Mařík, Miroslav Barta, and Matej Cibuľa. *Old Kingdom Art and Archaeology 7. Proceedings of the international conference. Universita degli Studi di Milano, 3–7 July 2017, pls. LXII–LXVIII [EDAL VI]*, chapter Invisible History: Hidden Markov Model of Old Kingdom administration development and its trends, pages 226–237. Milano: Pontremoli Editore, 2017.

[GIG17] 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.

[GLC+17] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang. Long text generation via adversarial training with leaked information, 2017.

[HHP+12] Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey, 2012. ELRA, European Language Resources Association.

[JCL+19] 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.

[KNT+18] Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *International Conference on Machine Learning*, pages 2611–2620. PMLR, 2018.

[KRL21] Y. Kirstain, O. Ram, and O. Levy. Coreference Resolution without Span Representations. *CoRR*, abs/2101.00434, 2021.

[LCG+20] 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.

[LHLZ17] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. *CoRR*, abs/1707.07045, 2017.

[LLW18] David Lowell, Zachary C Lipton, and Byron C Wallace. Practical obstacles to deploying active learning. *arXiv preprint arXiv:1807.04801*, 2018.

[LOG+19] 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.

[LP20] J. Luoma and S. Pyysalo. Exploring cross-sentence contexts for named entity recognition with bert, 2020.

[LPB16] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

[Luo05] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[Maa14] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[Mar11] Radek Marik. Emerging topic discovery, us 2011/0219001 a1, September 2011.

[Mar16a] Radek Marik. On design of data consistency verification. In Dusan Maga, Alexandr Stefek, and Tomas Brezina, editors, *Proceedings of the 2016 17th International Conference on Mechatronics – Mechatronika (ME), Prague, Czech Republic, December 7–9*. Czech Technical University in Prague, 2016.

[Mar16b] Radek Marik. On large genealogical graph layouts. In Broňa Brejová, editor, *Proceedings of the 16th ITAT Conference Information Technologies - Applications and Theory, WASACNA 2016 : Workshop on Algorithmic and Structural Aspects of Complex Networks and Applications, September 15-19, Tatranské Matliare, Slovakia*, pages 218–225, September 2016.

[Mar16c] Radek Marik. Tree-based genealogical graph layout. In Yifan Hu and Martin Nöllenburg, editors, *Graph Drawing and Network Visualization, 24th International Symposium, GD 2016, Athens, Greece, September 19-21*, volume ISBN: 978-3-319-50105-5 (Print) 978-3-319-50106-2 (Online), 2016.

[Mar17a] Radek Marik. *Efficient Genealogical Graph Layout*, pages 567–578. Springer International Publishing, Cham, 2017.

[Mar17b] Radek Marik. *On Multitree-Like Graph Layering*, pages 595–606. Springer International Publishing, Cham, 2017.

[Mar18a] Radek Marik. Feature space decomposition using information theory. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2018, pages 53:1–53:10, New York, NY, USA, 2018. ACM.

[Mar18b] Radek Marik. Layered graph force-driven vertex positioning. In *Proceedings of the 13th International In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) Funchal, Madeira, Portugal, 27-29 January, IVAPP 2018*, volume

3: IVAPP, pages 301–308, 2018.

[Mar19] Radek Marik. Multitree-like graph layering crossing optimization. In *Proceedings of 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Prague, Czech Republic, February 25-27*, volume 3: IVAPP, 2019.

[MBKK15] Radek Mařík, Pavel Bezpalec, Jan Kučerák, and Lukáš Kencl. Revealing viber communication patterns to assess protocol vulnerability. In *2015 International Conference on Computing and Network Communications (CoCoNet). Leonia, NJ 07605: EDAS Conference Services*, pages 502–510, 2015.

[MC17] Radek Marik and Matej Cibula. Multi-attribute sequence interpretation using hmm. In *The 2017 4th International Conference on Systems and Informatics (ICSAI 2017), 11-13 November, Hangzhou, China*, pages 1529–1534, 2017.

[MCCD13] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[MHK14] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[MHM18] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[MKP14] Radek Mařík, Zdeněk Kouba, and Michal Pantůček. Web application relations mapping. In Hendrik Decker, Lenka Lhotská, Sebastian Link, Marcus Spies, and RolandR. Wagner, editors, *Database and Expert Systems Applications*, volume 8645 of *Lecture Notes in Computer Science*, pages 155–163. Springer International Publishing, 2014.

[MZ19] Radek Marik and Tomas Zikmund. *Overlapping Communities in Bipartite Graphs*, volume 1, pages 207–218. Springer International Publishing, Cham, 2019.

[NDG+17] Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. Dynet: The dynamic neural network toolkit, 2017.

[NNP+21] Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, and Daniel Zeman. Coreference in universal dependencies 0.1 (CorefUD 0.1), 2021. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[Nov17] Michal Novák. Coreference resolution system not only for czech. In Jaroslava Hlavácová, editor, *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017), Martinské hole, Slovakia, September 22-26, 2017*, volume 1885 of *CEUR Workshop Proceedings*, pages 193–200. CEUR-WS.org, 2017.

[OFR+19] 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.

[RN18] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018.

[RWC+19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[Sah20] Marko Sahan. Active learning for text classification. Master's thesis, CTU in Prague, 2020.

[SMG19] 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.

[SSH19] J. Straková, M. Straka, and J. Hajič. Neural architectures for nested ner through linearization, 2019.

[SSM21] Marko Sahan, Vaclav Smidl, and Radek Marik. Active learning for text classification and fake news detection. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 87–94. IEEE, 2021.

[SYL+17] 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.

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

[VBA+95] Marc Vilain, John D Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[VSP+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017.

[WMWG17] 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.

[WPM+13] 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.

[WT11] 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.

[XSD20] P. Xia, J. Sedoc, and B. Van Durme. Incremental neural coreference resolution in constant memory, 2020.

[YAS+20] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention, 2020.

[YDY+20] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[Zik20] Tomáš Zikmund. Overlapping community detection in bipartite graphs. Master's thesis, CTU Prague, 2020.

[ZRZ+16] 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.

[ZW19] J. Zhanming and L. Wei. Dependency-guided lstm-crf for named entity recognition, 2019.