Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2019"

Moscow, May 29—June 1, 2019

SENTENCE LEVEL REPRESENTATION AND LANGUAGE MODELS IN THE TASK OF COREFERENCE RESOLUTION FOR RUSSIAN

Le T. A. (anhlt@vimaru.edu.vn)[1, 2],
Petrov M. A. (maksimallist@gmail.com)[1],
Kuratov Y. M. (yurakuratov@gmail.com)[1],
Burtsev M. S. (burtcev.ms@mipt.ru)[1]

[1]Neural Networks and Deep Learning Lab—Moscow Institute of Physics and Technology, Moscow, Russia

[2]Faculty of Information Technology—Vietnam Maritime University, Hai Phong, Viet Nam

Coreference Resolution (CR) is one of the most difficult tasks in the field of Natural Language Processing due to the lack of deeply and comprehensively understanding the semantic meaning of the mention in not only the sentence-level context but also the entire document-level context. To the best of our knowledge, the previous proposed models often address the coreference resolution task in two steps: 1) detect all possible mention candidates, 2) score and cluster them into chains. We instead propose a new approach which reforms the coreference resolution task to the task of learning sentence-level coreferential relations. Additionally, by leveraging the power of state-of-the-art language representation models such as BERT, ELMo, it was possible to achieve cutting edge results on Russian datasets.

Key words: coreference resolution, language modeling, sentence-level coreference

1. Introduction

Coreference resolution has a long research history but the quality of solutions is still not really convincing, especially for Russian language. As far as we know, there have been few deep learning-based coreference models that achieved state-of-the-art performances and all of them are studied on English datasets. Kevin et al. [6] proposed a variant of reinforcement learning solution with reward-rescaled max-margin objective to directly optimize a mention-ranking model for coreference evaluation metrics. This model obtained remarkable results on CoNLL2012 dataset [10], 65.73% and 63.88% on English and Chinese test sets, respectively. In the follow up paper [5] the problem was approached in a different way, where the entity-level information was captured with distributed representations of coreference cluster pairs. This model was trained with learning-to-search algorithm. The model performance was not better than the previous one.

Kenton Lee et al. [4], [3] proposed two end-to-end coreference models. The first one is the simple model with two steps: 1) create span representation from context-dependent boundary representations and head-finding attention mechanism, 2) cluster mentions. The second model is an improvement of the first one with inference procedure involving iterations of refining span representations. The model achieved 73% of average F1 on the test set of the English CoNLL-2012 shared task [10].

Starting with Kenton Lee et al. 's work as a baseline, this paper aims to build an end-to-end coreference model for Russian language. Previous works on coreference resolution on Russian language were mostly rule-based or feature-based simple models like random forest [15], [13]. The main contributions of this paper are listed bellow:

- Original model that learns to predict sentence-level coreferential relationships. The model is then directly integrated or generate features for the baseline model.
- Extention of the baseline model with state-of-the-art contextual language models ELMo and BERT trained for Russian language to boost task performance.

To test our proposed models, we participated in the shared task at Dialogue 2019 conference and achieved the best results in both tasks:

- Coreference task: The first place at the round using the gold mention boundaries, the second place at the round using only the raw text.
- Anaphora task: The first place at both rounds with or without using the gold mention boundaries.

2. Models

2.1. Baseline model

In this section we briefly describe the Higher-order Coreference Resolution model, which is used as baseline model (refer to the original papers [3] and [4] for more details).

Coreference resolution task consists of two sub-tasks: mention detection and mention clustering. This model solves both of them in end-to-end manner. Firstly, each text span is encoded by a single vector g_i, which is a concatenation of the first, last and head tokens representations. Secondly, mention score $s_m(i)$ is computed as $s_m(i) = w_{\text{m}} \cdot \text{FFNN}_m(g_i)$, where FFNN is a feed-forward neural network. Then top K (K depends on text length) spans are selected based on mention score. Finally, antecedent score $s_a(i, j)$ is computed for selected top K spans. Antecedent score should be positive if mention j is an antecedent to mention i. Then coreference chains are collected according to antecedent scores.

2.2. Sentence-level Coreferential Relationship-based Model

One of the main difficulties of coreference resolution task in comparison to other NLP tasks is the length of input text. Input of Name Entity Recognition task is one sentence. Question Answering models use several sentences as a context and one as a question. Meanwhile, input of coreference resolution task is one paragraph, or even one document with several hundred sentences. In order to make predictions correctly, the model need to capture the sentence semantics in the document context. Encoding a sentence in the context of document with several hundred sentences is a long-standing challenge. This challenge leads to the difficulty of applying common deep neural network models. In order to address this problem, we propose Sentence-level Coreferential Relationship-based model (SCRb model) that takes as input a document and outputs a square matrix representing the probabilities of coreferential links of sentences. In the training set this matrix is a binary square matrix (See **Fig. 1**, **2** for more details). The matrix is then can be used in two ways. In the first one, the probabilities produced by SCRb model are utilized as an input features. In the second way, the SCRb model is directly integrated into the end to end coreference model and both of them are trained jointly.

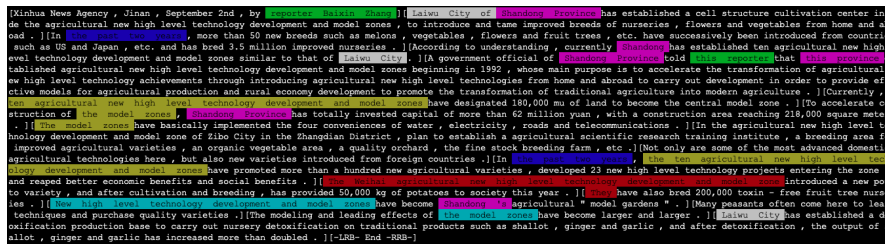

Figure 1: Visualization of the document chtb_0219.v4_gold_conll in the OntoNotes 5.0 dataset. The mentions with the same highlight color are belong to the same cluster

Figure 2: Binary matrix for document chtb_0219.v4_gold_conll representing the sentence-level coreferential relationship that SCRb model learns to predict

Here we describe step by step how SCRb model works:

- The model uses two types of word embedding: 1) free-context word embedding (e_{fc}) and 2) context-based word embedding (e_{cb}). In addition, to represent OOV words better, a convolutional network is utilized to generate character-based word embedding (e_{ch}). All these vectors are then concatenated to create the final word embedding:
$$e_w = [e_{fc}, e_{cb}, e_{ch}] \quad (1),$$
here [,] denotes the concatenation operator.
- The final word embeddings of each sentence are then feed into a Bi-LSTM network to output word vectors representing words in their sentence context:
$$w = \left[o_{\overrightarrow{lstm}}, o_{\overleftarrow{lstm}}\right], \quad (2)$$
here $o_{\overrightarrow{lstm}}$ and $o_{\overleftarrow{lstm}}$ are outputs of forward and backward LSTM networks, respectively.
- A maxpooling layer is used to reduce the word dimension to create the sentence representation:
$$s = max\_pooling(w_i), \quad (3)$$
where $w_i \in s$.
- The second Bi-LSTM network is utilized to capture the final sentence representation in the document context:
$$s_{dc} = \left[s_{\overrightarrow{lstm}}, s_{\overleftarrow{lstm}}\right] \quad (4)$$
- To create the matrix representing sentence relations, we modified Multi-dimensional Self-attention [12]:
  - Let $s_i \in \mathbb{R}^{d_s}$, where d_s denotes the length of sentence vectors outputted by the last Bi-LSTM network, is the vector representing the i^{th} sentence in the document.
  - Let $e_{d_{ij}} \in \mathbb{R}^{d_d}$, where d_d denotes the length of position encoding vectors, is distance embedding between s_i and s_j.

- Let $W \in \mathbb{R}^{d_s}$, $W_1, W_2 \in \mathbb{R}^{d_s \times d_s}, W_d \in \mathbb{R}^{d_s \times d_d}$ are weight matrices, and $b_1 \in \mathbb{R}^{d_s}, b \in \mathbb{R}$ are bias terms.
- The aligment score between s_i and s_j are computed as following formula:
$$f(s_i,s_j) = W^T\sigma(W_1 s_i + W_2 s_j + W_d e_{d_{ij}} + b_1) + b, \quad (5)$$
where σ is the activation function.
- Final antecedent score is computed as sum of $f(s_{mi}, s_{mj})$ and antecedent score $s_a(i,j)$ of baseline model, s_{m_i}—sentence, which mention i belongs to.

The graphical illustration of SCRb model is shown in Fig. 3.

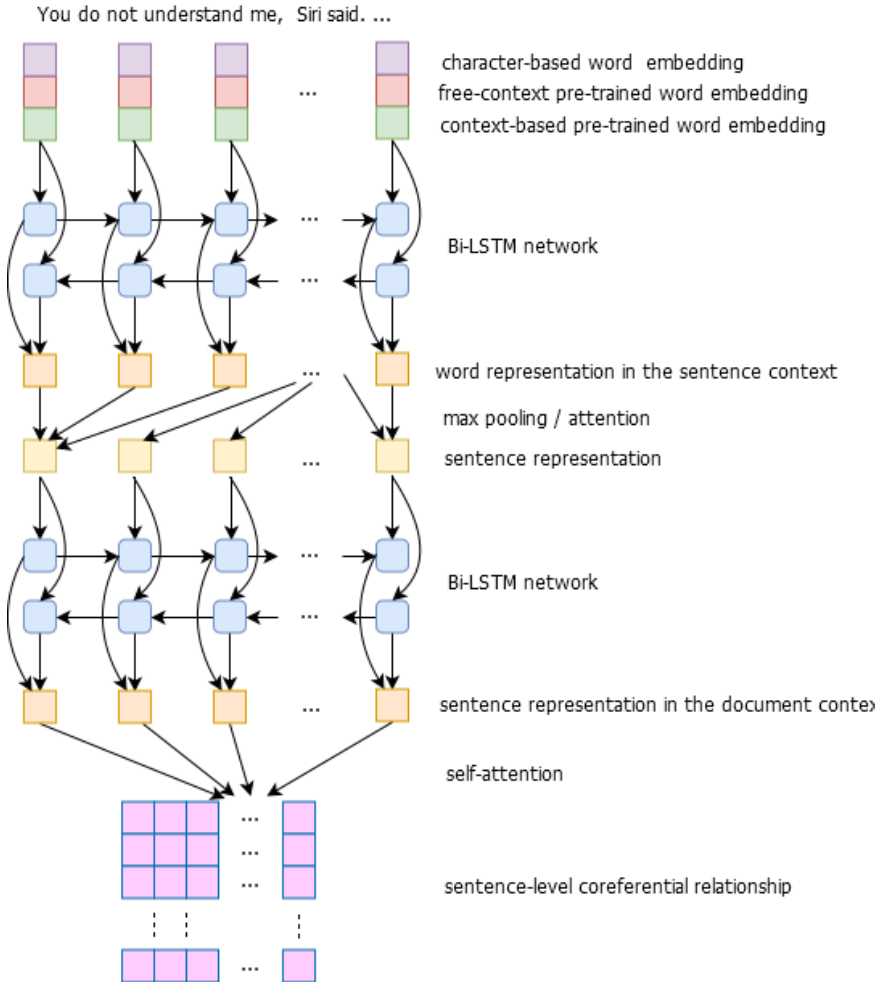

Figure 3: Sentence-level coreferential realtionship-based model

2.3. Model based on Language Modeling

Pretrained language models, such as ELMo [9], GPT [11], BERT [2], showed to be very effective in wide range of tasks from text classification to question answering. ELMo has been already tested on the task of coreference resolution for English language and helped to achieve new state-of-the-art performance [3] on CoNLL-2012 shared task dataset. Pretrained language models are usually used as a provider of contextualized word embeddings instead of usual word embeddings like w2v [8]. Contextualized word embeddings can be computed as weighted sum of outputs from each layer of language model, weights in this sum are trainable parameters, e.g., one scalar variable for word embedding layer and two scalars for BiLSTM layers are trained for ELMo, three parameters in total. BERT-base model is a 12-layer Transformer network and we did experiments with 1–6–12 and 10–11–12 layers outputs. The Higher-order Coreference Resolution model [3] uses two types of embeddings: word embeddings and contextualized word embeddings. We experimented with contextualized word embeddings from ELMo and BERT models trained for Russian Language (RuBERT).

3. Experiments and Results

3.1. Evaluation Metrics and Datasets

We did our experiments with three datasets, one for English language—CoNLL 2012 Shared Task[1] [10] and two datasets for Russian language: RuCor [15] from Dialogue-21 2014 Shared Task[2] and AnCor from Dialogue-21 2019 Shared Task.[3] As shown in **Table 1**, Russian datasets are 7–10 times smaller than the English one. This makes the CR task for Russian even harder.

Table 1: Coreference resolution datasets. Mentions and chains number computed for train + dev + test sets

Datasets	Language	Mentions	Chains
CoNLL 2012 Shared Task [10]	En	194,480	44,221
RuCor [15]	Ru	16,558	3,638
AnCor	Ru	28,961	5,678

There are three most common metrics for coreference resolution: MUC, B-cube, CEAF [7]. Overall coreference resolution systems performance is usually computed as averaged F-1 measure of these three metrics.

[1] http://conll.cemantix.org/2012/data.html

[2] http://www.dialog-21.ru/evaluation/2014/anaphora/

[3] http://www.dialog-21.ru/evaluation/

3.2. Experiments details

We used TensorFlow[4] to implement all models in our experiments. We took ELMo and RuBERT[5] models for Russian Language from DeepPavlov library[1]. For experiments with Russian language we used only raw texts without any additional features (like speaker id, morphological tags, etc) or pre-processing steps. All we have to do is to transform mention clusters in the original datasets to binary matrices representing the sentence relationships (as shown in **Fig. 2**).

All experiments were run on GeForce GTX 1080 Ti. The average training time is about one day for Russian datasets and one day and a half for the English one.

3.3. Results

In the first batch of experiments, information about the sentence-level coreferential relationship was supposed to be known before. In other words, we want to evaluate how sentence-level coreferential relationship affects the model performance. To do this, the baseline model is trained on two kinds of datasets: 1) the original OntoNotes 5.0; 2) the OntoNotes 5.0 with sentence-level coreferential relationships. The OntoNotes 5.0 with coreference chains was released as a part of CoNLL 2012 Shared Task. The experiment results pointed out that the information about the sentence relationships is a very useful feature for the coreference resolution task. If this feature is provided with 91% of accuracy the model performance can be boosted by about 2.5%. Under the ideal condition, when training with groundtruth sentence-level coreferential relationship, the model performance can be as large as 78.84% (refer **Table 2** for more details).

Table 2: Effect of sentence-level coreferential relationship on the baseline model performance

Dataset	Max. F1 on the dev. set
Original OntoNotes 5.0	73.00
OntoNotes 5.0 + sent.-level coref. relationship with 20% of noise	74.13
OntoNotes 5.0 + sent.-level coref. relationship with 16% of noise	74.73
OntoNotes 5.0 + sent.-level coref. relationship with 9% of noise	75.56
OntoNotes 5.0 + sent.-level coref. relationship with 6% of noise	76.36
OntoNotes 5.0 + sent.-level coref. relationship with 3.5% of noise	77.01
OntoNotes 5.0 + sent.-level coref. relationship with 1.5% of noise	77.92
OntoNotes 5.0 + groundtruth sent.-level coref. relationship	78.84

Results on AnCor and RuCor datasets were obtained by averaging results across 10-folds. RuBERT(1–6–12) with features from 1–6–12 layers showed better performance than RuBERT(10–11–12) in our preliminary experiments. In some experiments

[4] https://www.tensorflow.org/

[5] http://docs.deeppavlov.ai/en/master/components/bert.html

on AnCor dataset we also used RuCor dataset as additional training data (+ RuCor in Tables 5 and 6). Sentence-level information showed to be useful on RuCor dataset, it outperforms baseline model with about 1 F-1 point.

We tested our models in two settings:

- Gold mentions—uses gold mention boundaries and builds coreference chains (Tables 3 & 5).
- Full pipeline—includes mentions extraction from texts and building coreference chains (Tables 4 & 6).

Table 3: Results on RuCor dataset, gold mentions

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Sysoev [13]	69.28	63.12	55.33	62.58
Toldova [15]	70.25	60.14	—	—
Baseline + ELMo	90.54	79.71	67.81	79.36

Table 4: Results on RuCor dataset, full pipeline

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Sysoev [13]	41.90	34.30	29.06	35.10
Baseline + ELMo	67.26	52.29	53.18	57.58
SCRb	66.32	54.09	54.86	58.42

Table 5: Results on AnCor dataset, gold mentions

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Baseline + ELMo	90.22	83.41	59.44	77.69
Baseline + RuBERT(1–6–12)	91.04	84.38	63.07	79.50
Baseline + ELMo + RuCor	91.51	84.16	61.33	79.01
Baseline + RuBERT(1–6–12) + RuCor	91.47	84.49	63.81	79.92

Table 6: Results on AnCor dataset, full pipeline

Model	*muc*	*bcube*	*ceafe*	*avg.* F_1
Baseline + ELMo	50.29	48.89	46.99	51.72
SCRb	60.00	48.89	50.39	53.61
Baseline + RuBERT(1–6–12)	60.95	51.08	49.24	53.76
Baseline + ELMo + RuCor	65.01	52.67	50.19	55.96
Baseline + RuBERT(1–6–12) + RuCor	66.74	54.88	51.72	57.78

4. Discussions and Conclusions

In this paper, we presented a new approach to the task of coreference resolution with focus on Russian language. The previous models often address coreference resolution task in two stages: 1) detect all mention candidates, 2) cluster them into chains. We instead build a model to extract the sentence relations in the coreference context. This idea stems from an attempt of achieving sentence-level coreferential relationships to deal with the long term dependency. However, so far, the performance of the SCRb model is not really impressive. By analyzing the weights of the trained model, we found that the combined model tends to ignore the features learned by SCRb model. Hence, we claim that a part of the reason may lie in the way we combine the SCRb model with the baseline model. One more reason is the class imbalance problem that occurs when transforming mention clusters from original datasets to binary matrices. Although we used a weighted loss function that gives more importance to the minority classes, the problem has not been solved thoroughly. However, this model still has promising potentials to be applied to not only the CR task but also other NLP tasks such as Question Answering as well as Text Summarization. The experiment mentioned in the beginning of the **Section 3** shows that if the quality of the sentence-level coreferential feature is good enough, it can significantly boost the model performance.

In conclusion, we propose a new model that is able to learn the sentence-level coreferential relationships. In addition, we used two cutting edge language representation models (ELMo, BERT) to boost our model for Russian language. Our experiments and results on the shared tasks at Dialogue conference showed that our model achieved state-of-the-art performance on Russian language.

Acknowledgements

This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

References

1. *Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al.* Deeppavlov: Open-source library for dialogue systems. In Proceedings of ACL 2018, System Demonstrations, pages 122–127, 2018.
2. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
3. *Luheng He Kenton Lee and Luke Zettlemoyer.* Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of NAACL-HLT 2018, pages 687–692, 2018.

4. *Mike Lewis Kenton Lee, Luheng He and Luke Zettlemoyer.* End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, 2017.
5. *Chirstopher D.* Manning Kenvin Clark. Improving coreference resolution by learning entity-level distributed representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 643–653, 2016.
6. *Christopher D. Manning Kevin Clark*. Deep reinforcement learning for mention-ranking coreference models. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2256–2262, 2016.
7. *Xiaoqiang Luo.* On coreference resolution performance metrics. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pages 25–32. Association for Computational Linguistics, 2005.
8. *Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean*. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119, 2013.
9. *Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer.* Deep contextualized word representations. In Proc. of NAACL, 2018.
10. *Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang.* Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In Joint Conference on EMNLP and CoNLL-Shared Task, pages 1–40. Association for Computational Linguistics, 2012.
11. *Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever.* Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI, 2018.
12. *Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang.* Disan: Directional self-attention network for rnn/cnn-free language understanding. CoRR, abs/1709.04696, 2017.
13. *A. A. Sysoev, I. A. Andrianov, and Khadzhiiskaia A. Y.* Coreference resolution in russian: State-of-the-art approaches application and evolvement. In Computational Linguistics and Intellectual Technologies. International Conference "Dialogue 2017" Proceedings, pages 317–338, 2017.
14. *S. Toldova, A. Roytberg, A. A. Ladygina, M. D. Vasilyeva, I. L. Azerkovich, M. Kurzukov, G. Sim, D. V. Gorshkov, A. Ivanova, A. Nedoluzhko, and Y. Grishina*. Evaluating anaphora and coreference resolution for russian. In Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii Dialog, pages 681–695, 2014.
15. *Svetlana Toldova and Ionov Maxim.* Coreference resolution for russian: The impact of semantic features. In Computational Linguistics and Intellectual Technologies. International Conference" Dialogue 2017" Proceedings, pages 339–349, 2017.