RuCoCo: a new Russian corpus with coreference annotation

Vladimir Dobrovolskii
ABBYY

v.dobrovolskii@abbyy.com

Mariia Michurina
MIPT, RSUH
Moscow, Russia

marimitchurina@gmail.com

Alexandra Ivoylova
MIPT, RSUH
Moscow, Russia
a.m.ivoylova@gmail.com

Abstract

We present a new corpus with coreference annotation, Russian Coreference Corpus (RuCoCo). The goal of RuCoCo is to obtain a large number of annotated texts while maintaining high inter-annotator agreement. RuCoCo contains news texts in Russian, part of which were annotated from scratch, and for the rest the machine-generated annotations were refined by human annotators. The size of our corpus is one million words and around 150,000 mentions. We make the corpus publicly available[1].

Keywords: coreference corpus, coreference resolution, anaphora resolution, corpus annotation, Russian language

RuCoCo: новый русскоязычный корпус кореференции

Добровольский В.А.
ABBYY

v.dobrovolskii@abbyy.com

Мичурина М.А.
МФТИ, РГГУ
Москва, Россия

marimitchurina@gmail.com

Ивойлова А.М.
МФТИ, РГГУ
Москва, Россия
a.m.ivoylova@gmail.com

Аннотация

В этой статье мы представляем новый корпус кореференции для русского языка RuCoCo. Цель корпуса RuCoCo - получить большое количество размеченных текстов и одновременно с этим добиться высокого уровня согласия между аннотаторами. RuCoCo состоит из текстов новостей на русском языке, часть из которых была аннотирована с нуля, а для остальных текстов была выполнена машинная разметка и доработана аннотаторами-носителями языка. Размер нашего корпуса составляет один миллион слов и около 150 000 упоминаний. Корпус находится в открытом доступе.

Ключевые слова: корпус кореференции, разрешение кореференции, разрешение анафоры, создание корпуса, русский язык

1 Introduction

The task of coreference resolution was introduced at the Sixth Message Understanding Conference (Grishman and Sundheim, 1996), where the first dataset for coreference resolution task was introduced. The dataset consisted of 25 articles from Wall Street Journal (30,000 words). The annotation scheme

---

[1]https://github.com/vdobrovolskii/rucoco

was considered a standard until the release of ACE 2005 Multilingual Training Corpus for the 2005 Automatic Content Extraction (ACE) technology evaluation (Doddington et al., 2004). The corpus included texts in English, Chinese and Arabic and contained around 650,000 words in total for the three languages.

The MUC guidelines were domain-oriented, and their definition of a *markable* (mention) was mostly syntactically motivated. But further developments in this area, starting with the ACE initiative, increasingly involved semantic factors, so that recent corpora with coreference annotation define markables based on semantic class restrictions.

Quite a lot of such corpora were created in the last two decades, their primary goals being to increase the size in order to satisfy the requirements of the data-driven approach and to improve inter-annotator agreement which in many cases is too low, especially when a dataset addresses more complex cases of coreference.

The most well-known corpus of this kind is OntoNotes 5.0 (Pradhan et al., 2013). OntoNotes contains texts of various genres in three languages: English, Arabic, and Chinese. The cumulative volume of this corpus is 2.9 million words (about 1.5 million being English). The average annotator agreement for OntoNotes is 91.8% for normal coreference and 94.2% for appositives (Hovy et al., 2006).

The authors of the ARRAU corpus (Poesio et al., 2008; Uryupina et al., 2020) concentrate on "difficult" cases of anaphora: plural anaphora, abstract object anaphora, and ambiguous anaphoric expressions, so the corpus has bridging reference and discourse deixis annotated. It contains only English texts (although there is an Italian analogue LiveMemories (Rodríguez et al., 2010)); its current size is 350,000 tokens. The inter-annotator agreement in ARRAU varies from 67% (annotation of anaphoric ambiguities) to 95% (annotation of complex anaphoric relations).

Thus, most of the largest corpora with coreference annotation contain predominantly English texts; however, with the growing interest in natural language processing of Non-English languages, corpora in other languages are being developed more often. As for the Russian language, there now exist two such datasets, one of them being RuCor (Toldova et al., 2014; Toldova et al., 2015) and the other AnCor (Budnikov et al., 2019).

RuCor contains texts from openly available sources, such as Russian OpenCorpora, Lib.ru and Lenta.ru (156,000 words in total). In this corpus the annotation process was conducted over morphosyntactically pre-processed texts. The annotation scheme differentiates between primary and secondary markables, according to Potsdam Coreference Scheme (Krasavina and Chiarcos, 2007), where the primary markables are always annotated and represent specific references, while the secondary markables are annotated only if they are antecedents of any of the primary markables. Inter-annotator agreement for RuCor is 66% (Cohen's Kappa) or 85% (Mitkov's metric).

AnCor was created for the Ru-Eval competition in 2019 and contained 523 texts of various genres from Russian OpenCorpora (193,000 words in total). Named entities, common NPs and pronouns were annotated; the inter-annotator agreement for this dataset is 62.7% (75.5% agreement of both annotators and the final version).

As can be seen, although there are plenty of different corpora with coreference annotation, the largest and the most complex ones do not contain texts in Russian, and as for the Russian corpora, they are significantly smaller than the English ones, besides, their inter-annotator agreement is lower.

Therefore our main goals were to create a sufficiently big Russian corpus which would contain annotation of at least some difficult cases of anaphora with the inter-annotator agreement being high enough compared to OntoNotes and ARRAU.

2 RuCoCo: Russian Coreference Corpus

2.1 Data

We utilize the news stories published by NEWSru.com[2] as our source of text data. The texts were automatically collected and processed in the following way:

---

[2]https://www.newsru.com/

1. Any texts containing videos or embedded widgets from other websites were discarded as well as any texts marked as promotions.
2. Then the texts were converted to plain text format and cleared of any remaining HTML artifacts.
3. Texts that contained fewer than 20 tokens were also discarded, because they mostly consisted of a heading and a follow-up link only.
4. We then uniformly sampled one million words worth of texts across all text lengths and news categories. The total number of sampled texts is 3075.
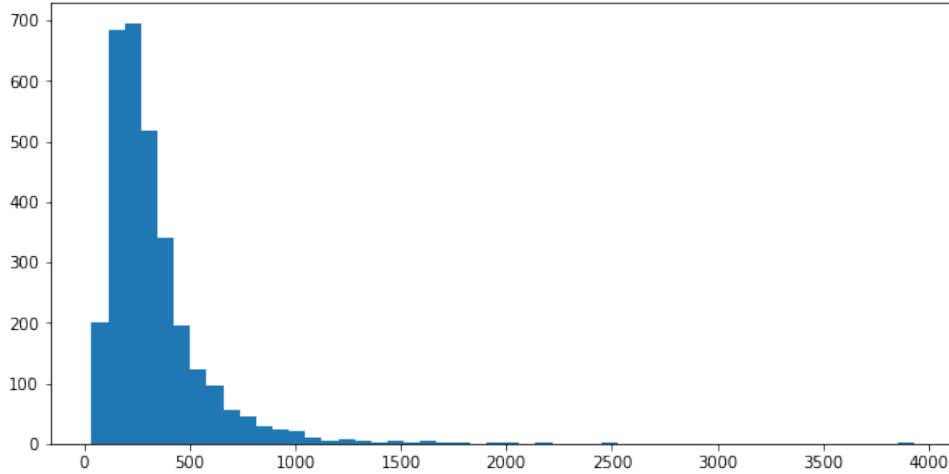
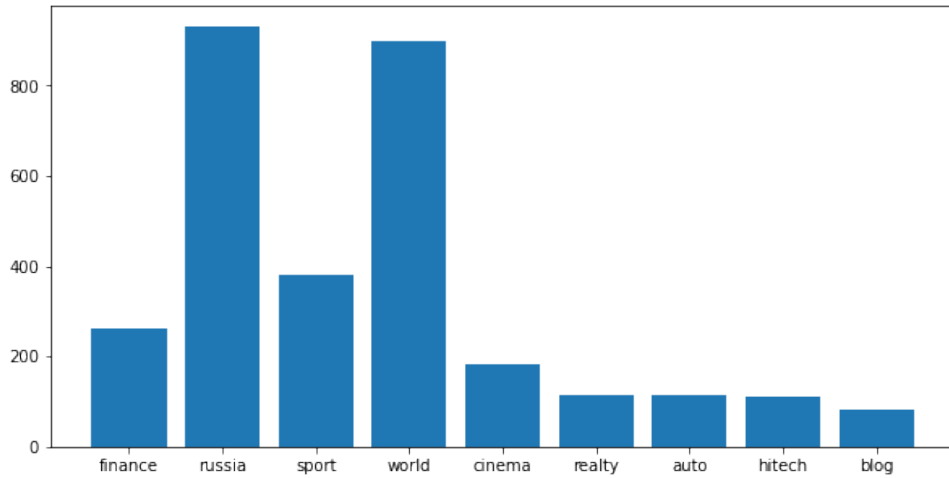Figure 1: Distribution of text lengths in the sampled data.

Figure 2: Distribution of news categories in the sampled data.

2.2 Annotation layer

The first release of RuCoCo covers identity (and in some cases, near-identity) coreference of noun phrases and pronouns. We do not annotate singletons, which means that each mention is linked to at least one other mention. We do not assign any attributes to the markables.

Mentions: We treat all noun phrases as potential mentions. Additionally the following types of pronouns are annotated:
- personal, possessive and reflexive pronouns;
- reciprocal pronouns, such as *друг друга* ('each other');
- relative pronouns;
- interrogative pronouns.

However, at this point we do not annotate coreference links with adjectives, clauses and expressions of time, all of which are going to be treated as valid mentions in the second revision of the corpus.

Mention boundaries: In most cases full noun phrases are annotated. To avoid overlapping of mentions referring to the same entity, participle and relative clauses that depend on the mention head are not included in mention boundaries. Therefore, in the following example there is no overlapping: *{клиент}_0, {который}_0 хотел пополнить {свой}_0 счет* ('{a customer}_0, {who}_0 wanted to top up {their}_0 account'). Parenthesis is not annotated unless it contains an independent clause, in which case it is treated as a regular sentence.

Coreference and anaphora: Coreference is annotated only for mentions of concrete entities. For generic mentions and mentions of abstract entities, events and properties we only annotate anaphora: *Может ли машина действовать разумно? Может ли {машина}_0 обладать сознанием? Может ли {она}_0 чувствовать?* ('Can a machine act intelligently? Can {a machine}_0 have a consciousness? Can {it}_0 feel how things are?'). Here, the first mention of *машина* ('a machine') is not annotated as coreferent with other mentions, because it is a generic mention.

Ellipsis: Mentions with elided heads are not annotated, as it would create ambiguity: *Это твоя сестра или {Даниэля}_0? Это {сестра {Даниэля}_0}_1, {она}_1 приехала на выходные.* ('Is this your sister or {Daniel's}_0? This is {{Daniel's}_0 sister}_1, {she}_1 came for the weekend.'). In the example above, the underlined mention could be recovered as *сестра Даниэля* ('Daniels' sister'), but we do not annotate it as referring to entity #1, because there would be two identical mentions referring to different entities.

Split antecedents: In RuCoCo, we annotate split antecedents as a means of dealing with the following challenges:

- Mentions referring to multiple referents: *{Премьер-министр}_0 и {госпожа Саймондс}_1 поженились вчера днем, небольшая церемония прошла в Вестминстерском соборе. {Пара}_{0,1} отпразднует свадьбу с семьей и друзьями следующим летом.*[3] ('{Prime Minister}_0 has married {Carrie Symonds}_1 yesterday afternoon in a "small ceremony" at Westminster Cathedral. {The couple}_{0,1} would celebrate again with family and friends next summer.').

- Coordinate dependents: *{Сборные России и Канады}_{0,1} ранее ни разу не встречались в финалах чемпионатов мира. <..> {Отечественные хоккеисты}_0 победили {канадцев}_1 со счетом 5-3 в Стокгольме в 1989 году.* ('{National teams of Russia and Canada}_{0,1} have not played in IIHF finals before. <..> {The Russian team}_0 defeated {the Canadians}_1 5-3 in Stockholm in 1989.')

Further in the text we refer to mentions linked to split antecedents as *plural anaphors* and to entities built from such mentions as *plural anaphor entities*. The number of such entities in the corpus can be seen in Table 1.

Category	Words	Mentions	Entities	PA-Entities	APA-Entities
russia	352,672	55,338	13,891	1,083 (7.8%)	2,471 (17.8%)
world	311,445	50,283	12,660	1,045 (8.3%)	2,122 (16.8%)
finance	94,015	11,739	3,020	176 (5.8%)	447 (14.8%)
sport	80,352	11,807	3,331	279 (8.4%)	705 (21.2%)
cinema	53,645	8,003	2,116	167 (7.9%)	431 (20.4%)
realty	34,227	4,509	1,274	72 (5.7%)	184 (14.4%)
hitech	31,365	3,895	1,080	77 (7.1%)	150 (13.9%)
auto	24,735	2,914	881	40 (4.5%)	94 (10.7%)
blog	17,649	1,917	624	39 (6.3%)	94 (15.0%)
Total	1,000,105	150,405	38,877	2,978 (7.7%)	6,698 (17.2%)

Table 1: Number of words, extracted mentions, entities, plural-anaphor (PA) entities and antecedent-of-plural-anaphor (APA) entitities across the news categories in RuCoCo.

[3] https://www.newsru.com/world/30may2021/bjohnson.html

Metonymy: Linking of metonymies is allowed: *{Лондон}$_0$ и {Брюссель}$_1$ официально объявили о соглашении по Brexit. {Евросоюзу}$_1$ и {Великобритании}$_0$ удалось выработать соглашение об отношениях после Brexit.* ('{London}$_0$ and {Brussels}$_1$ have announced a Brexit trade deal. {The European Union}$_1$ and {the United Kingdom}$_0$ have agreed on a post-Brexit trade deal.').

Corpus format: RuCoCo is distributed as a collection of JSON-formatted files. An entity is represented as a list of character offset pairs. Antecedents of plural anaphor entities are listed in the "includes" section.

```
{
    "entities" : [[[31, 34]], [[39, 42], [100, 103]], [[71, 75]]],
    "includes" : [[], [], [0, 1]],
    "text": "At half-past nine, that night, Tom and Sid were sent to bed, as
    ↪  usual. They said their prayers, and Sid was soon asleep.\n"
}
```

Listing 1: JSON-formatted annotation of the following example: *At half-past nine, that night, {Tom}$_0$ and {Sid}$_1$ were sent to bed, as usual. {They}$_{0,1}$ said their prayers, and {Sid}$_1$ was soon asleep.*

3 Corpus annotation

3.1 Metrics

There exist a number of coreference evaluation metrics, such as *MUC* (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), *CEAF* (Luo, 2005), *BLANC* (Recasens and Hovy, 2011) and others. Since the CoNLL-2012 shared task (Pradhan et al., 2012), the average score of *MUC*, B^3 and $CEAF_e$, has become a de-facto standard way to evaluate coreference resolution systems. However, several shortcomings of these three metrics were demonstrated by Moosavi and Strube (2016), who also introduced *LEA*, a coreference evaluation metric designed to overcome those shortcomings. *LEA* of a set of entities K is computed as:

$$\frac{\sum_{e_i \in E}(importance(e_i) \times resolutionScore(e_i))}{\sum_{e_j \in E} importance(e_j)} \tag{1}$$

where $importance(e) = |e|$ and the resolution score of entity k_i is calculated against the response set of entities R as follows:

$$resolutionScore(k) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)} \tag{2}$$

Here, $link(e)$ calculates the number of unique coreference link within e: $link(e) = |e| \times (|e| - 1)/2$.

We adopt *LEA* as our primary metric for measuring inter-annotator agreement and evaluating the neural coreference resolution model. As *LEA* does not support split antecedents out of the box, we modify the metric in the following way: for each plural anaphor entity we additionally calculate the scores of a special dummy entity with $importance$ set to be the number of antecedent entities and $resolutionScore$ computed based on the directed links between the plural anaphor entity and its antecedent entities.

The corpus was annotated by a team of 20 students of General Linguistics. The annotators were chosen based on trials that involved annotating documents of up to 1500 words. Each of the resulting documents was compared to the gold annotation using the *LEA* metric. The passing score was set to 0.9; the passing rate was 67%. Five of the annotators with the highest annotation quality were later appointed as moderators.

3.2 Neural pre-annotator

To speed up the annotation process, we developed a neural coreference resolution model to pre-annotate the texts. The model is based on the architecture proposed by Lee et al. (2018) and improved by Joshi et al. (2019) with the following differences:

- We use the Russian version of RoBERTa (Liu et al., 2019) pretrained by Sber AI[4].
- We replace the neural mention extraction module with a rule-based syntactic mention extractor built on top of spaCy (Honnibal and Montani, 2017). This allows us to explicitly define what a mention is instead of relying on neural networks for mention extraction.
- Following Dobrovolskii (2021), we represent mentions using only weighted sums of the subtoken embeddings that constitute the mention.

To train the model, we used the automatically merged annotations obtained during the early phases of annotation. We ignored plural anaphors and used the original *LEA* to evaluate the pre-annotation quality. The model performed at 0.62 F1 after being trained on 100,000 words, at 0.68 F1 after being trained on 400,000 words and at 0.73 F1 after training on the whole dataset of 1,000,000 words.

3.3 Annotation process

The annotation process consisted of two steps: the first 100,000 words were annotated from scratch, i.e. the task was to identify and link all coreferent mentions in raw texts; the remaining 900,000 words were first pre-annotated by a neural coreference resolution model and the annotators were asked to correct the resulting documents.

Each text in the corpus was annotated by two annotators and then finalized by a moderator who received an automatic merge of the two versions with differences highlighted. Additionally, 3500 words of each annotator were manually checked by the authors of the markup scheme to provide feedback on an early stage.

3.4 Inter-annotator agreement

We measured the inter-annotator agreement and found it to be 0.759 F1. Because the annotators do not have a closed set of mentions to link, we suspect that some of the differences between annotations can be attributed to lack of attention. To eliminate this factor, we conducted the following experiment on a subset of the data approximately 50,000 words in size: each annotator was given back their own annotations automatically merged with the other annotation versions. The annotators were asked to independently correct the documents. The resulting inter-annotator agreement was 0.890 F1.

3.5 Disagreement analysis

We analysed discrepancies of the two phases of corpus annotation: 1) from scratch (50 random texts, about 16,000 words examined) and 2) pre-tagged annotation (158 random texts, 50,000 words examined). Discrepancies were divided into several categories:

- missing/redundant coreference cluster;
- missing/redundant markable;
- missing/redundant anaphoric chain;
- plural anaphors with split antecedents;
- mentions referred to different entities;
- NP borders.

To make the comparison more informative, we carried out the error analysis of the neural model used for pre-tagging, although we need to keep in mind that after the first 100,000 words were checked, we made a number of minor clarifications and changes in the guidelines to facilitate the work of our annotation team. See the comparison of discrepancies in annotation from scratch, model errors and pre-tagged texts in Figure 3.

By **missing/redundant coreference cluster** we mean all cases when one of the two annotators skipped the whole cluster or marked up an unnecessary entity. It is the most frequent type when annotators
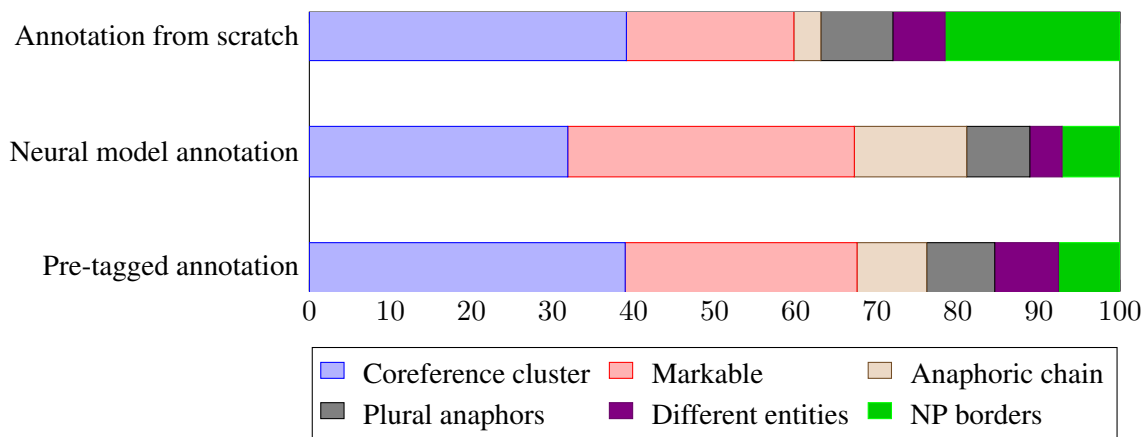
---

[4]https://github.com/sberbank-ai/model-zoo

Figure 3: Comparison of discrepancies on all annotation steps, %

disagree (about 39% for both annotation stages). There was no closed set of entities, moreover, for abstract and generic entities, events or referents denoting open sets (so-called non-concrete entities) only anaphora must be annotated. Thus, annotators should decide whether the entity is concrete or non-concrete. They disagree on the following examples: locations without proper names like *кризисный регион* ('crisis area'), *жилой квартал* ('residential area'), and also when locations are nested in an organisation name: *Россия* ('Russia') in *Министр транспорта России* ('Minister of Transport of Russia') or in *Верховном суде РФ* ('Supreme Court of the Russian Federation'). Some other popular types are events like *концерт в Москве* ('the concert in Moscow'), *чемпионат России по хоккею* ('Russian ice hockey championship') and some abstract entities that are very similar to events as they have participants like *контракт* ('contract'), *уголовное дело* ('criminal case').

Missing/redundant markable (about 20% and 28% respectively) is the case when an annotator missed one or several mentions, although the coreference cluster is there in both annotation versions. For these cases we examined types of NPs missed by one annotator in the annotation from scratch stage, having preserved the taxonomy as in (Toldova et al., 2015) in order to compare them. See Table 2 to check numbers. We can observe that both annotation groups of students tend to miss noun groups (i.e. noun phrases headed by a noun) more than any other NP type.

NP Type	Our Data, %	Toldova et al., 2015, %
Reflexive pronouns	4.73	3.76
Relative pronouns	1.77	6.20
Anaphoric pronouns	4.73	12.47
Possessive pronouns	2.37	6.48
Noun groups	85.2	71.08
Adverbs (here/there)	1.18	0.00

Table 2: Types of missed NPs

As for **missing/redundant anaphoric chain** (3.3% and 8.6%) i.e. chains with abstract or generic entities where only anaphora resolution was performed, annotators mostly missed chains containing a relative pronoun *который* ('which/that') as an anaphoric element e.g. *срок, до которого* ('the deadline by which'), *той политической линии, которую* ('the policy that').

In **plural anaphors with split antecedents** (9% out of all discrepancies, both stages), the most common discrepancy is a missing relation between a person and a group of people: a son and a family, *Кондолиза Райс* ('Condoleezza Rice'), *сенатор Хиллари Клинтон* ('Senator Hillary Clinton') and *политики* ('politicians'). Less frequent cases of disagreement are the following: part-whole relations (which are not annotated as split anaphora) and entities denoting several items with part of these items

as split antecedents: *50 терактов* ('50 terror attacks') and *20 терактов* ('20 terror attacks').

Mentions referred to different entities (6.4% and 7.9%) include cases where one or several mentions were assigned to different clusters by annotators in some confusing contexts (e.g. pronouns) or one annotator labelled some mentions in one and the same chain while the other one has divided it into several chains e.g. cases with metonymy like *Пхеньян* ('Pyongyang') and *КНДР* ('North Korea'), *Израиль* ('Israel') and *Израильская армия* ('Israel Defense Forces').

Disagreement on **NP borders** covers 21% of discrepancies in the first stage and substantially less on the pre-tagged stage (7.5 %). We may assume that it may be due to the ability of our model to find correct borders or that it is due to the clarified guidelines of syntactic ambiguities we made before the second annotation stage: we have highlighted that in all such cases the maximum NP border must be annotated. This category presupposes cases where annotators excluded modifiers as in *изменения* ('changes') vs. *самые существенные изменения* ('the most significant changes'), complements e.g. *Банк* ('the Bank') vs. *Банк России* ('the Bank of Russia') and less often appositives: *Берт Ньюборн* ('Burt Neuborne') vs. *Берт Ньюборн, профессор права Университета Нью-Йорка* ('Burt Neuborne Professor of Civil Liberties at New York University').

This analysis was presented to the moderators so that they would know what to pay attention to. Despite all these discrepancies, the resulting inter-annotator agreement is still 0.890 F1 and all the disagreements were resolved by our moderators.

4 Conclusion

The result of our work is the Russian Coreference Corpus, which is the largest corpus with coreference annotation for Russian so far. We managed to achieve almost 90% inter-annotator agreement; we also analyzed the most common disagreements between our annotators so that we know what issues are to be solved. Further developments will include annotating more difficult cases of anaphora as well as increasing the size and genre diversity of the corpus.

Acknowledgements

References

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

A. E. Budnikov, S. Yu. Toldova, D.S. Zvereva, D. M. Maximova, and M. I. Ionov. 2019. Ru-eval-2019: Evaluating anaphora and coreference resolution for russian. *// Computational Linguistics and Intellectual Technologies - Supplementary Volume*.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. *// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 7670–7675, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. *// Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. *// COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. // *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, P 57–60.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, P 5803–5808, Hong Kong, China, November. Association for Computational Linguistics.

Olga Krasavina and Christian Chiarcos. 2007. Pocos-potsdam coreference scheme. // *Proceedings of the Linguistic Annotation Workshop*, P 156–163.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, P 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

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

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 632–642, Berlin, Germany, August. Association for Computational Linguistics.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the arrau corpus. // *LREC*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. // *Joint Conference on EMNLP and CoNLL - Shared Task*, P 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. // *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, P 143–152, Sofia, Bulgaria, August. Association for Computational Linguistics.

M. Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17:485 – 510, 10.

Kepa Joseba Rodríguez, Francesca Delogu, Yannick Versley, Egon Stemle, and Massimo Poesio. 2010. Anaphoric annotation of wikipedia and blogs in the live memories corpus. // *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

S. Toldova, A. Roytberg, A. Ladygina, M. Vasilyeva, I. Azerkovich, M. Kurzukov, G. Sim, D. Gorshkov, A. Ivanova, A. Nedoluzhko, and Grishina Y. 2014. Ru-eval-2014: Evaluating anaphora and coreference resolution for russian. // *Computational linguistics and intellectual technologies: Proceedings of the international conference "Dialogue"*, P 681–694.

Svetlana Toldova, Ilya Azerkovich, Yulia Grishina, Alina Ladygina, Olga Lyashevkaya, Anna Roytberg, Galina Sim, and Maria Vasilieva. 2015. Pre-experiments on annotation of russian coreference corpus. *Higher School of Economics Research Paper No. WP BRP*, 35.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

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