УДК 004.85

ПОДХОД К СОЗДАНИЮ ДОМЕННО-ОРИЕНТИРОВАННОГО КОРПУСА ТЕКСТОВ С РАЗМЕТКОЙ КОРЕФЕРЕНЦИИ

Р. Д. Шувалов[1] [0009-0004-4056-0501], **Е. А. Сидорова**[2] [0000-0001-8731-3058]

[1]*Новосибирский государственный университет, г. Новосибирск,* Россия
[2]*Институт систем информатики им. А. П. Ершова СО РАН,*
г. Новосибирск, Россия

[1]r.shuvalov@g.nsu.ru, [2]lsidorova@iis.nsk.su

Аннотация

Представлен новый русскоязычный корпус с разметкой кореференции HaRuCo (Habr Russian Coreference Corpus). В качестве основы для корпуса взяты научно-популярные статьи, относящиеся к предметной области «Компьютерная лингвистика». Предложена методика разметки кореференции для текстов узких предметных областей, которая включает четыре основных этапа: синтаксический анализ текста; сборку именных групп и выделение местоимений для построения упоминаний (спанов); классификацию упоминаний классами предметной области; кластеризацию упоминаний в соответствии с цепочками кореферентно-связанных спанов. Аннотирование кореферентных связей осуществлено с применением синтаксического парсера и большой языковой модели, оно прошло ручную проверку и корректировку. Созданный корпус включает 3727 сущностей, 9905 упоминаний и 2683 кореферентных цепочек. Данный корпус может быть использован для обучения и оценки моделей разрешения кореференции для русскоязычных текстов ограниченной предметной области.

Ключевые слова*: разрешение кореференции, разметка кореференции, корпус текстов, научно-популярный текст, методика разметки, разметка упоминаний, кластеризация упоминаний, кореферентная связь.*

ВВЕДЕНИЕ

Кореференция (англ. coreference / co-reference) – это вид текстовой или синтаксической связности, при которой два или более фрагмента текста указы-

вают на один и тот же объект (референт) [1]. Эти фрагменты текста называют *упоминаниями*, а упоминаемый объект – *сущностью*. Обычно в тексте упоминания представлены именами собственными, именными группами или местоимениями. Задачей разрешения кореференции называют задачу выявления и группировки упоминаний, относящихся к одному и тому же объекту реального мира. Ее часто также рассматривают как задачу извлечения из текста всех референтных связей между упоминаниями с последующей группировкой на основе цепочек связанных упоминаний.

Разрешение кореференции представляет собой важную задачу, которая имеет решающее значение во многих приложениях обработки естественного языка, таких как извлечение и поиск информации, аннотирование текстов, ответы на вопросы, анализ настроений и машинный перевод. Существуют различные подходы к разрешению кореференции: на основе правил, методов машинного обучения, больших языковых моделей (Large Language Model, LLM). Однако большинство исследований проведено на англоязычном материале. Русскоязычные тексты обладают своей спецификой: развитой морфологией и большой грамматической вариативностью, что усложняет задачу разрешения кореференции. При этом для русского языка по-прежнему наблюдается дефицит данных для обучения и оценки работы моделей разрешения кореференции, особенно в специализированных предметных областях (ПО). По сравнению с небольшими текстами общей тематики, такими как новости или публицистические статьи [2], тексты научного или научно-популярного жанра, относящиеся к определенной ПО, в среднем имеют больший размер и оперируют специфичными сущностями, что значительно усложняет разрешение кореференции. Для таких текстов, как показано в работе [3], использование онтологической информации повышает качество решений.

В настоящей работе мы представляем новый русскоязычный корпус с разметкой кореференции HaRuCo (Habr Russian Coreference Corpus). В качестве основы для корпуса выбраны научно-популярные статьи с платформы Хабр (habr.com/ru), относящиеся к предметной области «Компьютерная лингвистика». Корпус содержит преимущественно длинные тексты: средняя длина текста составляет 2.6 тыс. слов или 146 предложений – и включает разметку сущностей ПО.

1. ОБЗОР ИССЛЕДОВАНИЙ, БЛИЗКИХ ПО ТЕМАТИКЕ

В 2014 г. на конференции «Диалог» состоялось соревнование по разрешению кореференции Ru-Eval-2014 [4]. Участникам был предоставлен корпус с морфологическими и кореференционными аннотациями RuCor, состоящий из 181 документа. Тексты содержали от 5 до 100 предложений, самый длинный из них насчитывал 170 предложений. В общей сложности вручную было размечено 2000 пар анафорических местоимений и антецедентов, а также 1200 кореферентных цепочек. Корпус RuCor содержит тексты из общедоступных источников, таких как российский OpenCorpora, Lib.ru и Lenta.ru (всего 156 тыс. слов). В этом корпусе аннотирование проводилось вручную на основе предварительно обработанных текстов с помощью морфосинтаксического анализатора. Всего корпус содержит 29 тыс. упоминаний и около 5.5 тыс. кореферентных цепочек.

Следующий корпус AnCor был создан для соревнования по разрешению анафоры и кореференции Ru-Eval-2019 [5] и содержит 523 текста различных жанров из российской OpenCorpora общей длиной 193 тыс. слов. Средняя длина текста составляет 369 слов. Тексты были аннотированы автоматически, проверены и скорректированы вручную. Корпус содержит 5.7 тыс. кореферентных цепочек и 25 тыс. упоминаний.

Самым большим корпусом с кореферентной разметкой для русского языка является корпус RuCoCo, представленный в [2]. Целью создания корпуса RuCoCo было 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 автоматическая разметка, которая была доработана аннотаторами-носителями языка. Размер корпуса составляет 1 млн слов и около 150 тыс. упоминаний и 38.8 тыс. сущностей. Количество слов в текстах в основном меньше 500.

Существующие русскоязычные корпуса кореференции содержат в основном короткие тексты, в среднем не превышающие 500 слов или 100 предложений. В этих корпусах содержатся тексты, которые в большинстве своем характеризуются универсальной лексикой и, как следствие, не позволяют оценить,

насколько хорошо предлагаемые методы будут работать в условиях ограниченной предметной области.

2. КОРПУС HARUCO

2.1. Данные

Для построения корпуса кореференции за основу был взят корпус аналитических статей по компьютерной лингвистике с сайта Хабр, размеченный сущностями ПО [6]. В корпус входит 20 достаточно длинных текстов общим объемом 51.8 тыс. слов (средняя длина текста – 2.6 тыс. слов / 146 предложений). Распределение длин текстов представлено на рис. 1а.

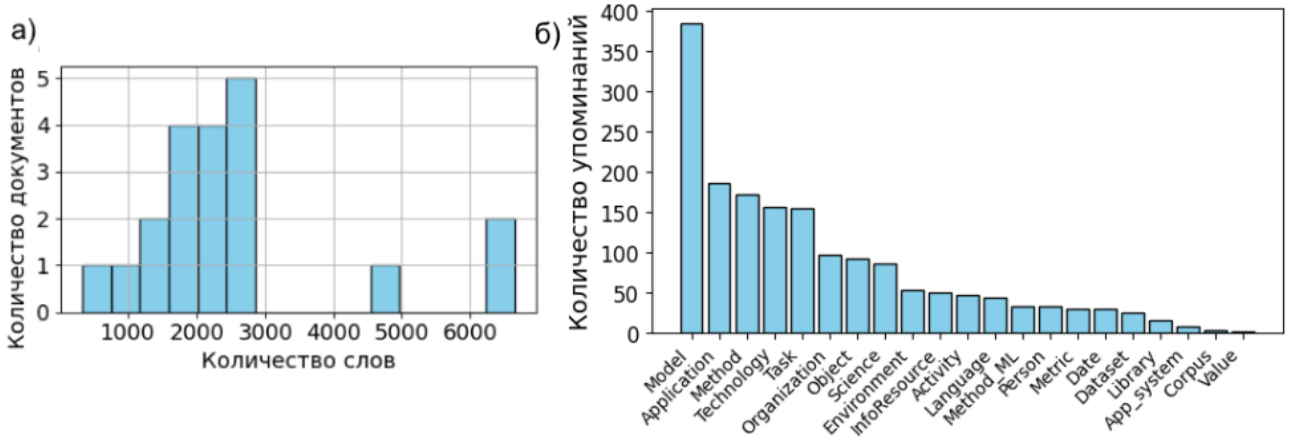

Рис. 1. Статистические характеристики базового корпуса: а) распределение длин текстов, б) распределение количества эталонных упоминаний по классам.

Начальная разметка корпуса включает разметку упоминаний предметных сущностей, представленных именными группами (без местоимений), относящихся к классам Model (Модель), Application (Приложение), Method (Метод), Technology (Технология), Organization (Организация) и др. (всего 21 понятие). На рис. 1б представлено распределение упоминаний по классам ПО.

Из представленной диаграммы видно, что наиболее распространенными являются упоминания специфичных для ПО сущностей (Модель, Приложение, Метод) в отличие от универсальных корпусов, в которых наиболее частотными являются упоминания Персон, Организаций и Локализации.

При дальнейшей работе эта разметка использовалась в качестве эталонных упоминаний за исключением упоминаний для классов Date (Дата) и Value (Значение) – для них разметка кореференции не проводилась.

2.2. Принципы разметки кореференции

Для обеспечения совместимости с самым большим набором русскоязычных данных RuCoCo для аннотирования была использована та же схема с небольшими модификациями. Разметка упоминаний в данной схеме охватывает имена собственные, именные группы и местоимения, для которых нашлось не менее одного связанного упоминания. Разметка кореференции включает разметку групп связанных упоминаний (кластеров) с учетом множественных упоминаний и возможных пересечений границ именных групп [2].

Внесенные изменения связаны с ориентацией используемого корпуса на решение задачи извлечения информации. Так, разметка в HaRuCo включает только упоминания сущностей, относящихся к области компьютерной лингвистики, в том числе одиночные упоминания (такие сущности образуют кластеры, состоящие из одного элемента). Кроме того, в отличие от RuCoCo, кореференция размечалась не только для конкретных упоминаний, но и для общих, или абстрактных, упоминаний, что связано с особенностями онтологического представления ПО.

2.3. Характеристики корпуса

В результате был создан новый аннотированный корпус со следующими характеристиками: общее число сущностей – 3727, количество упоминаний – 9905, количество кореферентных цепочек – 2683. На рис. 2 представлены распределения упоминаний по кластерам (объектам) и классам.

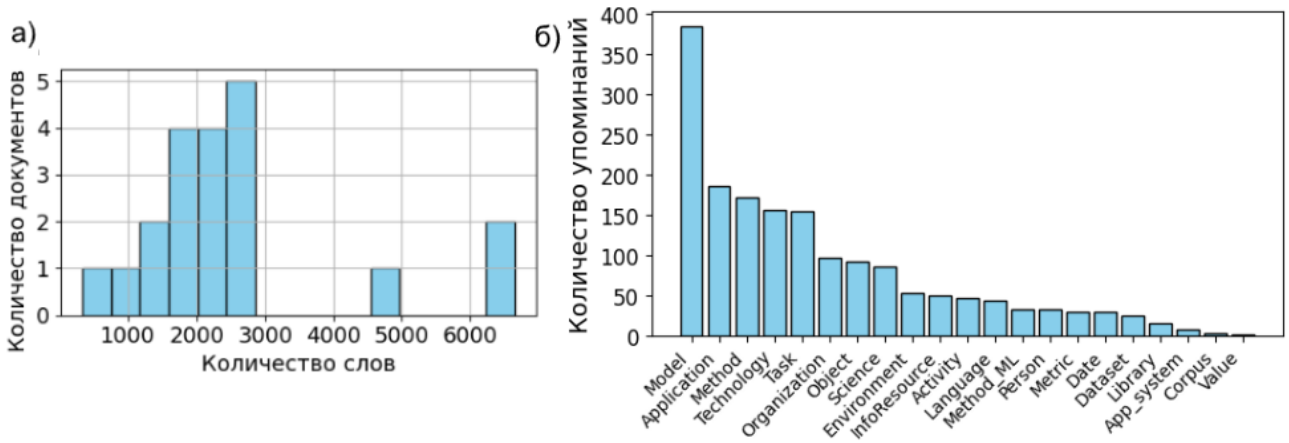

Рис. 2. Статистические характеристики корпуса HaRuCo: а) распределение упоминаний по кореферентным кластерам, б) распределение количества упоминаний сущностей по классам.

3. МЕТОДИКА АННОТИРОВАНИЯ

Разметка корпуса осуществлялась автоматически с последующей ручной корректировкой. Можно выделить четыре основных этапа процедуры разметки: синтаксический анализ текста; сборка именных групп и выделение местоимений для построения упоминаний (спанов); классификация упоминаний классами предметной области (на основе словаря и/или обученной языковой модели); кластеризация упоминаний в соответствии с цепочками кореферентно-связанных спанов.

Для выделения местоимений был использован синтаксический парсер, построенный на основе библиотеки SpaCy. Для связывания упоминаний был использован, представленный в [7] подход, в котором текст с размеченными упоминаниями подается на вход LLM вместе с промптом. LLM выделяет кластеры и присваивает идентификатор кластера каждому из размеченных упоминаний. Дополнительно в разметку текстов была добавлена нумерация упоминаний, что позволило впоследствии сопоставлять идентификаторы кластеров спанам.

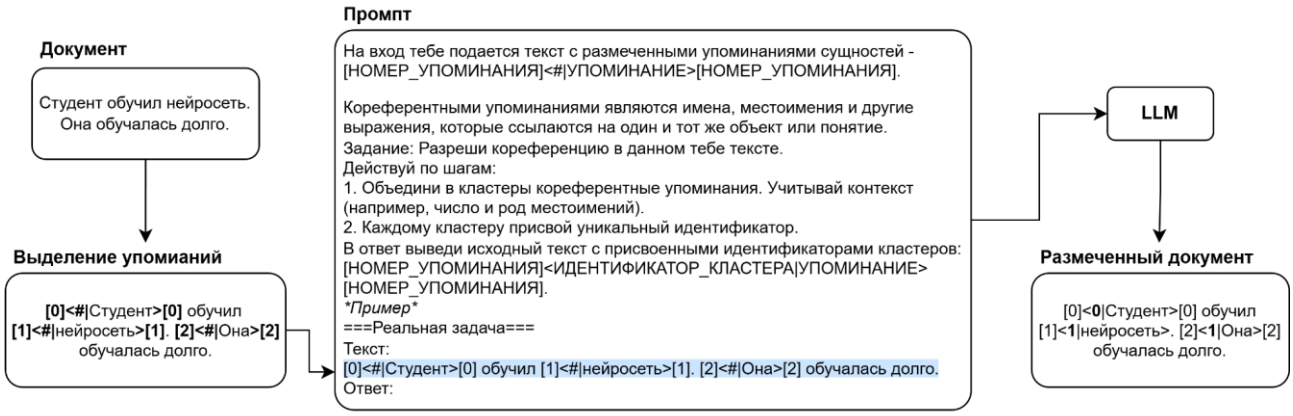

Рис. 3. Схема автоматической разметки кореференции.

В экспериментах были использованы модель Qwen2.5-32B-Instruct и промпт для русскоязычных текстов, представленный на рис. 3. В табл. 1 приведены оценки качества работы LLM для:

- связывания предварительно размеченных упоминаний;
- полной кореферентной разметки именных групп и местоимений.

Оценка осуществлялась на основе метрики LEA (Link-based Entity-Aware metric) [8].

Табл. 1. Результаты автоматической разметки.

Задача	Точность	Полнота	F1-мера
Связывание упоминаний	0.3	0.73	0.43
Разрешение кореференции	0.33	0.1	0.16

Из таблицы видно, что при использовании предварительно размеченных упоминаний модель показывает хорошую полноту, что позволило применять полученные результаты для дальнейшей ручной корректировки разметки. Результаты второго эксперимента показали, что модели данного класса пока плохо справляются с задачей разрешения кореференции в постановке end-to-end.

ЗАКЛЮЧЕНИЕ

Представлен русскоязычный корпус кореференции, включающий научно-популярные тексты, из предметной области «Компьютерная лингвистика». Разметка кореференции в корпусе совмещена с разметкой сущностей предметной

области. В корпус входят относительно длинные тексты, превышающие в среднем в 5–6 раз по длине большинство текстов из других русскоязычных корпусов. Методика разметки кореференции включает предварительный синтаксический анализ и выделение упоминаний, классификацию упоминаний, разметку кореференции с помощью LLM и дальнейшую ручную корректировку. Качество автоматической разметки зависит от моделей, используемых для выделения и связывания упоминаний, а также от качества инструкции, поэтому результаты для различных доменов могут отличаться. Созданный корпус может быть использован для обучения и оценки моделей разрешения кореференции для русского языка, а также для исследования кореференции в научно-популярных текстах на русском языке. Предложенная методика может быть использована для разметки кореференции в текстах узких предметных областей. Дальнейшая работа будет направлена на расширение данного корпуса, а также на адаптацию предложенного метода разрешения кореференции для обработки более длинных текстов, длина которых (с учетом промпта) превышает максимальный контекст для LLM.

СПИСОК ЛИТЕРАТУРЫ

1. *Падучева Е.В.* Высказывание и его соотнесенность с действительностью. М.: URSS, 2008.

2. *Dobrovolskii V.A., Michurina M.A., Ivoylova A.M.* RuCoCo: a new Russian corpus with coreference annotation // Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conference “Dialogue 2022”. 2022. P. 141–149. https://doi.org/10.28995/2075-7182-2022-21-141-149

3. *Azerkovich I.* Using Semantic Information for Coreference Resolution with Neural Networks in Russian // Analysis of Images, Social Networks and Texts. AIST 2019. Communications in Computer and Information Science. Cham: Springer International Publishing. 2020. Vol. 1086. P. 85–93. https://doi.org/10.1007/978-3-030-39575-9_9

4. *Toldova S. et al.* Ru-eval-2014: Evaluating anaphora and coreference resolution for russian // Computational linguistics and intellectual technologies: Proc. of the Int. Conference "Dialogue 2014". 2014. P. 681–694.

5. *Budnikov A.E., Toldova S.Yu., Zvereva D.S., Maximova D.M., Ionov M.I.*

Ru-eval-2019: Evaluating anaphora and coreference resolution for russian // Computational Linguistics and Intellectual Technologies – Supplementary Volume. 2019.

6. *Овчинникова К.А., Иванов А.И., Сидорова Е.А.* Автоматизация построения терминологического ядра онтологии по компьютерной лингвистике на основе корпуса текстов // Системная информатика. 2023. No. 23. С. 13–32.

7. *Nghia T. Le, Ritter A.* Are Large Language Models Robust Coreference Resolvers? // First Conference on Language Modeling (COLM-2024). 2024.

8. *Moosavi N.S., Strube M*. 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. 2016. Vol. 1. P. 632–642. https://doi.org/10.18653/v1/P16-1060

HARUCO: A NEW RUSSIAN-LANGUAGE CORPUS OF POPULAR SCIENCE TEXTS WITH COREFERENCE ANNOTATION

R. D. Shuvalov[1] [0009-0004-4056-0501], E. A. Sidorova[2] [0000-0001-8731-3058]

[1] *Novosibirsk State University, Novosibirsk, Russia*

[2] *A. P. Ershov Institute of Informatics Systems, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[1]r.shuvalov@g.nsu.ru, [2]lsidorova@iis.nsk.su

Abstract

This paper presents a new Russian-language corpus with coreference annotation, HaRuCo (Habr Russian Coreference Corpus). The corpus is based on popular science articles related to the subject area of “Computational Linguistics.” The paper proposes a method for coreference annotation in texts from narrow subject areas, which includes four main stages: syntactic analysis of the text, assembly of noun phrases and identification of pronouns for constructing mentions (spans), classification of mentions by subject area classes, and clustering of mentions according to chains of coreferentially linked spans. Coreferential links were annotated using a syntactic parser and a large language model, followed by manual verification and correc-

tion. The created corpus includes 3.727 entities, 9.905 mentions, and 2.683 coreferential chains. The created corpus can be used for training and evaluating coreference resolution models for the Russian language.

Keywords*: coreference resolution, coreference annotation, text corpus, popular science text, annotation methodology, mention annotation, mention clustering, coreference link.*

REFERENCES

1. *Paducheva E.V.* Vyskazyvanie i ego sootnesennost s deistvitelnostiu. M.: URSS, 2008.

2. *Dobrovolskii V.A., Michurina M.A., Ivoylova A.M.* RuCoCo: a new Russian corpus with coreference annotation // Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conference “Dialogue 2022”. 2022. P. 141–149. https://doi.org/10.28995/2075-7182-2022-21-141-149

3. *Azerkovich I*. Using Semantic Information for Coreference Resolution with Neural Networks in Russian // Analysis of Images, Social Networks and Texts. AIST 2019. Communications in Computer and Information Science. Cham: Springer International Publishing. 2020. Vol. 1086. P. 85–93. https://doi.org/10.1007/978-3-030-39575-9_9

4. *Toldova S. et al.* Ru-eval-2014: Evaluating anaphora and coreference resolution for russian // Computational linguistics and intellectual technologies: Proc. of the Int. Conference "Dialogue 2014". 2014. P. 681–694.

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

6. *Ovchinnikova K.A., Ivanov A.I., Sidorova E.A.* Automation of the construction of the terminological core of ontology in computer linguistics based on a corpus of texts // System Informatics. 2023. No. 23. P. 13–32. https://doi.org/10.31144/SI.2307-6410.2023.N23.P13-32

7. *Nghia T. Le, Ritter A.* Are Large Language Models Robust Coreference Resolvers? // First Conference on Language Modeling (COLM-2024). 2024.

8. *Moosavi N.S., Strube M.* 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. 2016. Vol. 1. P. 632–642. https://doi.org/10.18653/v1/P16-1060

СВЕДЕНИЯ ОБ АВТОРАХ

ШУВАЛОВ Роман Денисович – магистрант Факультета информационных технологий Новосибирского государственного университета. Научные интересы: программная инженерия, обработка естественного языка, технологии искусственного интеллекта.

Roman Denisovich SHUVALOV – graduate student at the Department, Information Technology, Novosibirsk State University. Research interests: software engineering, natural language processing, artificial intelligence technologies.

email: r.shuvalov@g.nsu.ru

ORCID: 0009-0004-4056-0501

СИДОРОВА Елена Анатольевна – кандидат физико-математических наук, старший научный сотрудник Института систем информатики им. А.П. Ершова СО РАН, доцент кафедры программирования и кафедры систем информатики Новосибирского государственного университета. В списке научных трудов более 160 работ в области компьютерной лингвистики, онтологического инжиниринга и разработки интеллектуальных систем.

Elena Anatolievna SIDOROVA – PhD (2006). She is a senior researcher at the A. P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, associate professor at Novosibirsk State University. List of scientific works includes more than 160 peer-reviewed publications in the fields of computational linguistics, intelligent system development, and knowledge and ontology engineering.

email: lsidorova@iis.nsk.su

ORCID: 0000-0001-8731-3058

Материал поступил в редакцию 12 апреля 2026 года