Создание русскоязычного корпуса аспектов аргументации

И. Н. Фищева, e-mail: fishchevain@gmail.com
Т. А. Пескишева, e-mail: peskisheva.t@mail.ru
В. С. Головизнина, e-mail: golovizninavs@gmail.com
Е. В. Котельников, e-mail: kotelnikov.ev@gmail.com

ФГБОУ ВО «Вятский государственный университет»

Аннотация. *Рассматривается процедура создания первого русскоязычного текстового корпуса с разметкой по аспектам аргументации. Под аспектом понимается слово или словосочетание, указывающее на одну из сторон или свойство целевого объекта, упоминаемого в утверждении, входящем в состав аргумента. Размеченный корпус позволит обучать нейросетевые модели классификации аргументов по аспектам, а также создавать модели генерации аргументативных текстов, содержащих заданные аспекты.*

Ключевые слова: *анализ аргументации, аспекты аргументов, глубокие нейросетевые модели.*

Введение

Анализ аргументации (argument mining) – это область компьютерной лингвистики, которая посвящена извлечению из текстов и классификации аргументов и связей между ними, а также построению аргументационной структуры [1].

Под аргументом понимается логическая структура, включающая в качестве ключевых компонентов основное утверждение и один или несколько доводов [2]. Основное утверждение отражает в текстовом виде определенную точку зрения (позицию) относительно целевого объекта. Доводы могут поддерживать или опровергать точку зрения, выраженную в основном утверждении. В первом случае довод отражает позицию «за» данную точку зрения, во втором случае – «против». При этом каждый довод описывает один или несколько аспектов целевого объекта [3]. Аспект определяется как слово или словосочетание, указывающее на одну из сторон или свойство целевого объекта, упоминаемого в заданном утверждении. Например, для целевого объекта «Электромобили» возможными аспектами могут быть «Удобство и комфорт», «Надежность», «Экологичность», «Стоимость».

Рассмотрим пример аргумента относительно целевого объекта «Электромобили» в соответствии с описанной терминологией. Пусть

основное утверждение имеет вид: «Электромобили лучше обычных авто». В этом утверждении выражена позиция «за» относительно целевого объекта. В качестве довода в поддержку данной позиции можно использовать высказывание «Стоимость батарей только за последние четыре года снизилась более чем в два раза» – в этом высказывании упоминается аспект «Стоимость». Пример довода «против»: «Отсутствие развитой сети электрозаправок является одной из причин медленного развития электромобилей в России» – здесь обсуждается аспект «Удобство и комфорт».

Обработка аргументации с учетом аспектов представляет собой более глубокий уровень анализа по сравнению с выявлением только доводов «за» и «против». Аспектно-ориентированный анализ аргументации позволяет осуществлять поиск, ранжирование, классификацию и генерацию аргументов в точном соответствии с потребностями пользователя [3, 4]. Соответствующие программные инструменты могут применяться в ходе деловых совещаний для убеждения оппонентов, в образовании при анализе и оценке аргументации в студенческих эссе, в научной работе в ходе анализа аргументов в заданной области исследований.

Однако, большинство работ в области извлечения аргументации проводятся на материале английского языка; исследований для русского за последние годы было немного [5–8]. Кроме того, насколько нам известно, работы в области аспектно-ориентированного анализа аргументации для русского языка отсутствуют.

В настоящей статье описывается создание первого русскоязычного текстового корпуса с разметкой по аспектам аргументации. Корпус предоставлен в общий доступ: https://github.com/kotelnikov-ev/RuArgumentMining.

1. Отбор потенциальных аргументов

Процесс создания корпуса включал два основных этапа – отбор предложений, потенциально содержащих аргументы, и разметка отобранных предложений.

С целью организации поиска качественных аргументов была создана база данных на основе информационно-поисковой системы Elasticsearch [9], включающая русскоязычную Википедию (3 994 609 документов) и корпус новостей Lenta.ru (https://github.com/yutkin/Lenta.Ru-News-Dataset) (800 976 документов).

Спектр тематик потенциальных аргументов был определен с использованием сайта ВЦИОМ (https://wciom.ru), на котором публикуются аналитические социологические обзоры по актуальным для современного российского общества темам. В результате был

сформирован перечень из 17 тематик (см. таблицу 1), таких как «Пенсионные сбережения», «Детские гаджеты» и «Донорство крови». Для каждой тематики был сформулирован дискуссионный вопрос (например, «Надо ли откладывать деньги на пенсию?») и утверждение, относительно которого осуществляется разметка аргументов (например, «Нужно делать пенсионные сбережения»).

С целью поиска потенциальных аргументов для каждой тематики был подготовлен список поисковых запросов. В результате выполнения данных поисковых запросов в информационно-поисковой системе Elasticsearch была собрана коллекция, включающая 520 документов. Дубликаты статей исключались. Собранные документы были подвергнуты сегментации на предложения при помощи библиотеки Razdel проекта Natasha (https://natasha.github.io/razdel).

Для автоматической классификации предложений на два класса – «аргумент» / «не аргумент» была обучена нейросетевая модель ArgBERT, основанная на предобученной модели sbert_large_mt_nlu_ru (https://huggingface.co/sberbank-ai/sbert_large_mt_nlu_ru). Обучение ArgBERT осуществлялось с использованием переводных версий англоязычных корпусов ArgMicro, Persuasive Essays и UKP Sentential Argument Mining Corpus [8]. Модель ArgBERT предоставлена в общий доступ: https://github.com/kotelnikov-ev/RuArgumentMining/tree/main/ArgBERT.

В результате классификации предложений было выделено в качестве потенциальных аргументов 15,4% от общего количества найденных предложений. Для разметки случайным образом были отобраны ровно 5000 потенциально аргументативных предложений с сохранением пропорции по тематикам, полученной исходно в ходе поиска.

2. Разметка аспектов

На следующем этапе осуществлялась разметка предложений по аргументам, которую выполняли три аннотатора. Разметка осуществлялась по трем параметрам: 1) является ли предложение аргументом по отношению к заданному утверждению; 2) если предложение является аргументом, то выражена позиция «за» или «против»; 3) если предложение является аргументом, то какой упоминается аспект. В качестве аргументационных аннотаторы рассматривали такие предложения, которые потенциально могут быть использованы для убеждения оппонента в дискуссии относительно данного утверждения.

Было выделено 20 универсальных аспектов, каждый из которых мог относиться к любой рассматриваемой тематике, например,

«Безопасность», «Влияние на здоровье», «Стоимость». Полный перечень аспектов приведен по ссылке: https://github.com/kotelnikov-ev/RuArgumentMining/tree/main/AspectCorpus.

Результатом разметки стал корпус, состоящий из 5000 предложений, 548 из которых размечены минимум двумя аннотаторами как аргументационные. Распределение предложений по тематикам приведено в таблице 1.

Таблица 1

Распределение аргументационных предложений по тематикам

Тематика	**Википедия**	**Lenta.ru**	**Всего**	
			Кол-во	**%**
Криптовалюта	145	53	198	36,1%
Детские гаджеты	50	24	74	13,5%
Электромобили	27	39	66	12,0%
Пенсионные сбережения	29	12	41	7,5%
Супермаркеты и продуктовые рынки	18	12	30	5,5%
Удаленная работа	12	15	27	4,9%
Покупки в интернете	11	9	20	3,6%
Донорство крови	18	0	18	3,3%
Детские видеоблоги	17	0	17	3,1%
Шутеры	0	15	15	2,7%
Бумажные и электронные книги	5	8	13	2,4%
Киберспорт	11	0	11	2,0%
Онлайн-образование	7	2	9	1,6%
Фриланс	8	1	9	1,6%
Детские лагеря	0	0	0	0,0%
Свободные деньги	0	0	0	0,0%
Электросамокаты	0	0	0	0,0%
Всего	**358**	**190**	**548**	**100,0%**

Согласие между аннотаторами, вычисленное по метрике Fleiss' kappa [10], составляет для корпуса Lenta.ru – 0,6458, для Википедии – 0,6249, в целом – 0,6427, что соответствует значительной (substantive) степени согласия [11].

Для одного аргументационного предложения аннотатор мог выделить от одного до трех аспектов. В результате для 548

аргументационных предложений было выделено 847 аспектов. Наибольшее количество аспектов отобрано для тематик «Криптовалюта» (327), «Электромобили» (106) и «Детские гаджеты» (92). Аспекты с максимальным количеством предложений: «Безопасность» (133 предложения), «Надежность» (90 предложений), «Удобство и комфорт» (88 предложений).

Заключение

Таким образом, основной результат работы – первый русскоязычный текстовый корпус с разметкой по аспектам аргументации, предоставленный в общий доступ. Корпус содержит 5000 предложений из русскоязычной Википедию и корпуса новостей Lenta.ru.

Размеченный корпус позволит обучать нейросетевые модели классификации аргументов по аспектам, а также создавать модели генерации аргументативных текстов, содержащих заданные аспекты.

Также обучена нейросетевая языковая модель ArgBERT, позволяющая осуществлять бинарную классификацию предложений на «аргумент» / «не аргумент».

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 22-21-00885, https://rscf.ru/project/22-21-00885/.

Литература

1. Lawrence, J. Argument Mining: A Survey / J. Lawrence, C. Reed // Computational Linguistics. – 2020. – Vol. 45(4). – P. 765–818.

2. Stede, M. Argumentation Mining. Synthesis Lectures on Human Language Technologies / M. Stede, J. Schneider. – Morgan & Claypool Publishers, 2018.

3. Schiller, B. Aspect-Controlled Neural Argument Generation / B. Schiller, J. Daxenberger, I. Gurevych // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2021. – P. 380–396.

4. Ruckdeschel, M. Boundary Detection and Categorization of Argument Aspects via Supervised Learning / M. Ruckdeschel. G. Wiedemann // Proceedings of the 9th Workshop on Argument Mining. – 2022. – P. 126–136.

5. Salomatina, N. Identification of connected arguments based on reasoning schemes “from expert opinion” / N. Salomatina, I. Kononenko, E. Sidorova, I. Pimenov // Journal of Physics: Conference Series. – 2021. – Vol. 1715.

6. Kononenko, I. The Study of Argumentative Relations in Popular Science Discourse / I. Kononenko, E. Sidorova, I. Akhmadeeva // RCAI 2020: Artificial Intelligence. – 2020. – P. 309–324.

7. Fishcheva, I. Cross-lingual argumentation mining for Russian texts / I. Fishcheva, E. Kotelnikov // Lecture Notes in Computer Science. – 2019. – Vol. 11832. – P. 134–144.

8. Fishcheva, I. Argumentative Text Generation in Economic Domain / I. Fishcheva, D. Osadchiy, K. Bochenina, E. Kotelnikov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue-2022". – 2022. – Issue 21. – P. 211–222.

9. Gormley, C. Elasticsearch: The definitive guide: A distributed real-time search and analytics engine / C. Gormley, Z. Tong. – O'Reilly Media Inc., 2015.

10. Fleiss, J.L. Measuring nominal scale agreement among many raters / J.L. Fleiss // Psychological Bulletin. – 1971. – Vol. 76(5). – P. 378–382.

11. Artstein, R. Inter-Coder Agreement for Computational Linguistics / R. Artstein, M. Poesio // Computational Linguistics. – 2008. – Vol. 43(4). – P. 555–596.