УДК 004.912 doi: 10.25730/VSU.0536.18.26

Извлечение аргументации из текстов и проблема отсутствия русскоязычных текстовых корпусов*

Е. В. Котельников
кандидат технических наук, доцент кафедры прикладной математики и информатики,
Вятский государственный университет. Россия, г. Киров. E-mail: ev_kotelnikov@vyatsu.ru

Аннотация. В статье рассматривается одно из перспективных направлений в современной компьютерной лингвистике – извлечение аргументации из текстов (Argumentation Mining). Перечисляются задачи, решаемые в системах извлечения аргументации, указываются области применения таких систем. Приводятся схема представления аргументации на основе теории Фримена и пример разметки текста с использованием данной схемы. Рассматриваются существующие текстовые корпуса, снабженные разметкой в соответствии с некоторой схемой аргументации. Отсутствие подобных русскоязычных корпусов является существенным препятствием для развития области Argumentation Mining в России. Два способа получения таких корпусов – на основе разметки новых текстов и с помощью профессионального перевода существующих корпусов на русский язык – оказываются весьма трудоемкими. Предлагается для формирования корпусов использовать машинный перевод и обозначается план дальнейшего исследования этой проблемы.

Ключевые слова: аргументация, извлечение аргументации из текстов, текстовые корпуса, машинный перевод.

1. Введение

В современной компьютерной лингвистике одним из наиболее интересных и перспективных направлений является *извлечение аргументации из текстов* (Argumentation Mining) [5–7]. Под *аргументацией* понимается вербальная, социальная и рациональная активность, направленная на убеждение разумного критика в приемлемости определенной точки зрения за счет выдвижения группы высказываний, подтверждающих или опровергающих данную точку зрения [15]. Предметом области Argumentation Mining является автоматическое обнаружение аргументов, представленных в тексте, связей между ними и структуры каждого отдельного аргумента [8].

Модули автоматического извлечения аргументации могут применяться для расширения возможностей систем анализа мнений в контексте выявления причин возникновения тех или иных мнений; для идентификации обоснования решений в юридических документах при поиске прецедентов; для раскрытия структуры аргументации в учебных работах с целью предоставления обратной связи студентам [1; 10].

При разработке модуля автоматического извлечения аргументации из текстов необходимо решить следующие задачи [9; 10]:

1) выявить фрагменты текста, содержащие аргументацию;
2) осуществить сегментацию найденных фрагментов на отдельные элементы, называемые аргументативными дискурсивными единицами (argumentative discourse units, ADU);
3) классифицировать ADU в соответствии с используемой схемой представления аргументации;
4) установить наличие и вид связей между всеми парами ADU.

Важную роль при решении указанных задач играет выбор схемы представления аргументации (или аргументационной схемы), в основу которой положена некоторая теория аргументации. Одна из наиболее влиятельных теорий была разработана С. Тулмином [14]. Он ввел шесть аргументационных ролей высказываний: «утверждение» (conclusion), «данные» (data), «основания» (warrant), «поддержка» (backing), «опровержение» (rebuttal), «определитель» (qualifier). Теория С. Тулмина была пересмотрена Дж. Фрименом, который ввел макроструктуру аргументов, позволяющую объединять высказывания, играющие различные роли, в аргументационную схему, отражающую процесс аргументации [4].

А. Пельдцусом и М. Штеде была предложена схема представления аргументации на основе теории Фримена [10]. В этой схеме *аргументом* называется совокупность посылок (premises), под-

** Работа выполнена при поддержке Deutscher Akademischer Austauschdienst (DAAD) и Министерства образования и науки Российской Федерации в рамках государственного задания Минобрнауки РФ № 2.12728.2018/12.2 по теме «Проведение научно-исследовательских работ в рамках международного научно-образовательного сотрудничества по программе "Михаил Ломоносов" по теме: "Разработка и исследование аннотированного русскоязычного текстового корпуса для анализа аргументации"».*

держивающих некоторый вывод или заключение (claim, conclusion). Посылки и заключения являются ADU. Для представления компонентов аргументов и связей между ними используется ряд графических обозначений, основные из которых представлены на рис. 1.

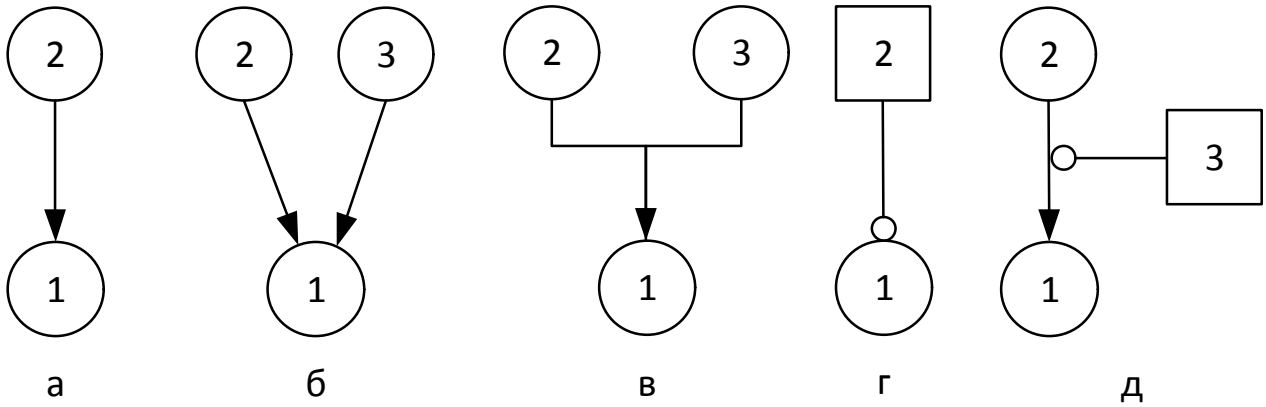

Рис. 1. Графические обозначения, используемые для представления аргументов

На рис. 1а обозначено заключение (1), поддерживаемое посылкой (2), – это простейший случай связи двух ADU (basic argument). Две независимые посылки (2) и (3), поддерживающие одно и то же заключение (1), показаны на рис. 1б (multiple support). Каждая из них может использоваться отдельно от другой – они никак не связаны, в отличие от случая на рис. 1в, где представлены связанные посылки (2) и (3) – одна дополняет другую (linked support).

На рис. 1г показано ADU (2), атакующее заключение (1) (rebut a conclusion). Другой вариант контраргумента представлен на рис. 1д, где ADU (3) атакует не само заключение (1), а связь между заключением (1) и посылкой (2) (undercut an argument).

Рассмотрим пример. Пусть дан следующий текст: «Пенсионный возраст должен быть повышен. Из-за низкой рождаемости доля пожилого населения и расходы на пенсионную систему возрастают. Да, рабочая нагрузка увеличивается во многих профессиях, но люди, становясь старше, остаются здоровыми благодаря современной медицине».

Данный текст можно представить в виде аргументационной схемы, показанной на рис. 2.

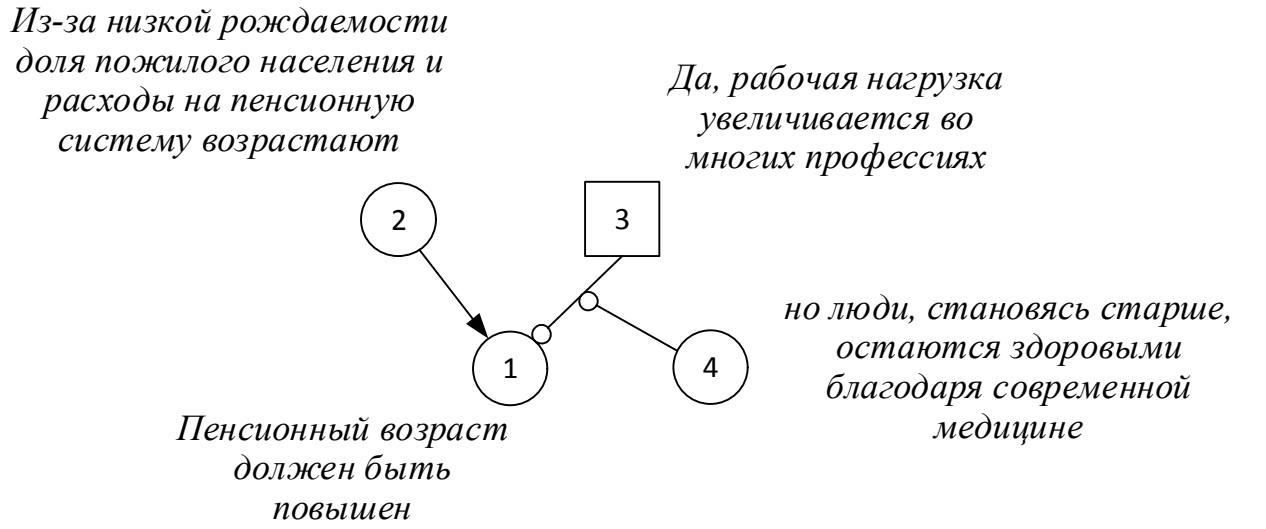

Рис. 2. Пример аргументационной схемы

На рис. 2 показаны заключение (1), посылка (2) в поддержку заключения, посылка (3), направленная против заключения, и посылка (4), атакующая связь посылки (3) и заключения (1).

Для построения систем извлечения аргументации из текстов необходимы текстовые корпуса, снабженные разметкой в соответствии с некоторой аргументационной схемой, например представленной на рис. 1 и 2. Для английского языка существует несколько подобных корпусов (схемы разметки для них отличаются); в качестве примеров можно привести следующие:

1) корпус аргументационных микротекстов[1] (Argumentative Microtext Corpus) [9] – это 112 коротких текстов (пять-шесть предложений) на разные тематики (повышение пенсионного возраста, медицинское страхование, школьная форма и т. п.) на английском и немецком языках, размеченные в соответствии со схемой, предложенной в [10];

2) корпус убеждающих эссе[2] (Persuasive Essays) [12] – это 402 эссе на английском языке (в среднем 17 предложений на эссе), обосновывающих определенные точки зрения по широкому спектру тематик – миграция, образование, бюджетная политика и т. д.

На веб-сайте AIFdb Corpora[3], поддерживаемом университетом Данди (Шотландия), собрано более 150 текстовых корпусов разного объема с разметкой аргументации. При этом в настоящее время не существует русскоязычных корпусов, размеченных в соответствии с какой-либо аргументационной схемой. Этот факт является существенным препятствием для развития области Argumentation Mining в России.

Есть два возможных способа формирования таких корпусов:

1) поиск текстов, содержащих аргументацию, и их разметка в соответствии с выбранной схемой;

2) перевод существующих корпусов на русский язык.

Оба способа являются весьма трудоемкими. В первом сначала необходимо найти тексты с аргументами. Возможными источниками могут служить научные статьи, политические дискурсы, судебные решения, редакционные статьи, студенческие эссе. Далее следует разметить (аннотировать) найденные тексты в соответствии с выбранной схемой аргументации. Для этого необходимо задействовать нескольких аннотаторов, тщательно объяснить им правила разметки, предоставить удобный инструмент разметки (для этого можно использовать GraPAT [11] или Brat [13]). По окончании процесса разметки следует проверить согласие между аннотаторами, например, на основе метрики каппы Флейса [3]. В случае высокой степени несогласия желательно согласовать мнения аннотаторов, возможно, повторив процедуру разметки для некоторых текстов.

Другой способ подразумевает качественный перевод существующих корпусов на русский язык (как правило, с английского). Трудоемкость этого способа также высока: в работе [2] указано, что перевод корпуса убеждающих эссе (7 141 предложение) [12] с английского языка на немецкий с сохранением аргументационной разметки занял 270 часов и стоил 3 000 долларов.

Поэтому предлагается исследовать вариант автоматического перевода размеченных корпусов. С этой целью для корпуса аргументационных микротекстов [9] было получено три варианта машинного перевода: с помощью систем Google Translate[4], Яндекс.Переводчик[5] и Promt[6], а также перевод профессионального переводчика.

В дальнейших исследованиях предполагается проанализировать качество машинного перевода на основе сравнения результатов извлечения аргументации для вариантов русскоязычных корпусов, созданных автоматически и профессиональным переводчиком.

Список литературы

1. *Afantenos S., Peldszus A., Stede M.* Comparing decoding mechanisms for parsing argumentative structures // Argument & Computation. 2018. Preprint, pp. 1–16.

2. *Eger S., Daxenberger J., Stab C., Gurevych I.* Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need! // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 831–844.

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

4. *Freeman J. B.* Argument Structure: Representation and Theory // Argumentation Library. 2011. Vol. 18. Springer.

5. *Habernal I., Gurevych I.* Argumentation Mining in User-Generated Web Discourse // Computational Linguistics. 2017. Vol. 43(1), pp. 125–179.

6. *Lippi M., Torroni P.* Argumentation Mining: State of the Art and Emerging Trends // ACM Transactions on Internet Technology. 2016. Vol. 16(2), pp. 1–25.

7. *Moens M.-F.* Argumentation mining: How can a machine acquire common sense and world knowledge? // Argument & Computation. 2018. Vol. 9, pp. 1–14.

8. *Palau R. M., Moens M.-F.* Argumentation mining: the detection, classification and structure of arguments in text // Proceedings of the 12th international conference on artificial intelligence and law. ACM. 2009, pp. 98–107.

9. *Peldszus A., Stede M.* An annotated corpus of argumentative microtexts // Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015. Vol. 2. London. College Publications, 2015, pp. 801–816.

[1] *http://angcl.ling.uni-potsdam.de/resources/argmicro.html.*

[2] *https://www.informatik.tu-darmstadt.de/ukp/research_6/data/index.en.jsp*

[3] *http://corpora.aifdb.org*

[4] *https://translate.google.com.*

[5] *https://translate.yandex.ru.*

[6] *https://www.translate.ru.*

10. *Peldszus A., Stede M.* From Argument Diagrams to Argumentation Mining in Texts: A Survey // International Journal of Cognitive Informatics and Natural Intelligence (IJCINI). 2013. Vol. 7(1), pp. 1–31.

11. *Sonntag J., Stede M.* GraPAT: a tool for graph annotations // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. Reykjavik, Iceland.

12. *Stab C., Gurevych I.* Parsing Argumentation Structure in Persuasive Essays // Computational Linguistics. 2017. Vol. 43(3), pp. 619–659.

13. *Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J.* Brat: a Web-based Tool for NLP-Assisted Text Annotation // Proceedings of the Demonstrations Session at EACL. 2012.

14. *Toulmin S.* The Uses of Argument. Cambridge University Press, Cambridge, 1958.

15. *Van Eemeren F.H., Grootendorst R., Johnson R.H., Plantin C., Willard Ch.A.* Fundamentals of argumentation theory: Handbook of historical background and contemporary developments. Routledge, 1996.

Extraction of argumentation from texts and the problem of the lack of Russian-language text corpora

E. V. Kotelnikov

PhD of technical sciences, associate professor of the Department of applied mathematics and informatics, Vyatka State University. Russia, Kirov. E-mail: ev_kotelnikov@vyatsu.ru

Abstract. One of the most promising directions in modern computational linguistics – text argumentation mining – is considered in the article. The tasks of argumentation mining systems are enumerated. The fields of application of such systems are indicated. An argumentation scheme based on Freeman's theory and an example of text annotation with the use of this scheme are given. Existing text corpora annotated in accordance with certain scheme are considered. The lack of Russian annotated text corpora is an essential obstacle for the development of argumentation mining in Russia. There are two ways of creating of such corpora – based on the annotation of new texts and with the help of professional translation of existing corpora. But both of them turn out to be very laborious. The use of machine translation for this task is proposed. Also the plan of further research in this direction is suggested.

Keywords: argumentation, text argumentation mining, text corpora, machine translation.

References

1. *Afantenos S., Peldszus A., Stede M.* Comparing decoding mechanisms for parsing argumentative structures // Argument & Computation. 2018. Preprint, pp. 1–16.

2. *Eger S., Daxenberger J., Stab C., Gurevych I.* Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need! // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA, August 20-26, 2018, pp. 831–844.

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

4. *Freeman J. B.* Argument Structure: Representation and Theory // Argumentation Library. 2011. Vol. 18. Springer.

5. *Habernal I., Gurevych I.* Argumentation Mining in User-Generated Web Discourse // Computational Linguistics. 2017. Vol. 43(1), pp. 125–179.

6. *Lippi M., Torroni P.* Argumentation Mining: State of the Art and Emerging Trends // ACM Transactions on Internet Technology. 2016. Vol. 16(2), pp. 1–25.

7. *Moens M.-F.* Argumentation mining: How can a machine acquire common sense and world knowledge? // Argument & Computation. 2018. Vol. 9, pp. 1–14.

8. *Palau R. M., Moens M.-F.* Argumentation mining: the detection, classification and structure of arguments in text // Proceedings of the 12th international conference on artificial intelligence and law. ACM. 2009, pp. 98–107.

9. *Peldszus A., Stede M.* An annotated corpus of argumentative microtexts // Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015. Vol. 2. London. College Publications, 2015, pp. 801–816.

10. *Peldszus A., Stede M.* From Argument Diagrams to Argumentation Mining in Texts: A Survey // International Journal of Cognitive Informatics and Natural Intelligence (IJCINI). 2013. Vol. 7(1), pp. 1–31.

11. *Sonntag J., Stede M.* GraPAT: a tool for graph annotations // Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). 2014. Reykjavik, Iceland.

12. *Stab C., Gurevych I.* Parsing Argumentation Structure in Persuasive Essays // Computational Linguistics. 2017. Vol. 43(3), pp. 619–659.

13. *Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J.* Brat: a Web-based Tool for NLP-Assisted Text Annotation // Proceedings of the Demonstrations Session at EACL. 2012.

14. *Toulmin S.* The Uses of Argument. Cambridge University Press, Cambridge, 1958.

15. *Van Eemeren F.H., Grootendorst R., Johnson R.H., Plantin C., Willard Ch.A.* Fundamentals of argumentation theory: Handbook of historical background and contemporary developments. Routledge, 1996.