***И.С. Пименов***

**Специфика аргументационного аннотирования научных и научно-популярных текстов**

**Аннотация.** Рассматривается задача ручной разметки аргументации в научных и научно-популярных русскоязычных текстах. Метод построения графов аргументации реализуется через набор веб-инструментов, обеспечивающих создание тематических корпусов, визуализацию аргументативных утверждений и схем аргументации. Указываются специфические сложности в применении метода к текстам выбранных жанров. Приводятся сравнительные количественные характеристики аргументационных структур текстов разных жанров и тематик.

**Ключевые слова.** Аргументационное аннотирование, схемы Уолтона, граф аргументации, научные и научно-популярные тексты.

**1. Введение**

Особой задачей в области автоматического извлечения информации выступает распознавание в тексте аргументов, или выстраиваемых рассуждений, с определением организующих их отношений. Изучение аргументации (способов её выражения в естественном языке, методов её извлечения) актуально в свете следующих приложений:

1. анализ отзывов (для определения не только тональности текста, но и обоснования оценки);
2. понимание и ведение дебатов;
3. принятие решений в рекомендательных системах;
4. обнаружение радикальных мнений, вводящих в заблуждение текстов и т.д.

Задача аннотирования аргументации отличается сложностью анализируемых сущностей (пропозиций, выражаемых в сегментах текста различной длины при значительном разнообразии лексических и грамматических средств) и связей, в качестве которых выступают модели рассуждений. Близким к аргументационной разметке по сложности является аннотирование риторических структур [Pisarevskaya et al. 2017].

В данной работе описывается эксперимент по ручной разметке аргументации в научных и научно-популярных текстах. Жанровая специфика выбранного материала характеризуется значительным объёмом текстов (в противовес новостным сообщениям, интернет-комментариям, эссе и микротекстам, из которых чаще всего составляются аргументационные корпуса [Daxenberger et al. 2017]), а также особенностями применяемых схем (в том числе ввиду монологической организации текстов). Помимо этого, перспектива машинной обработки аннотаций и использования коллекции для машинного обучения накладывает определённые ограничения на действия аннотатора.

Обзор литературы показывает, что корпусов научных текстов с аргументационной разметкой немного: в [Lauscher et al. 2018] описывается создание такого корпуса на материале английского языка с указанием на отсутствие аналогов.

Цель работы: адаптация методики аргументационного аннотирования к разметке научных и научно-популярных текстов с учетом применения аннотаций для машинного обучения.

**2. Задача аргументационного аннотирования**

Аргументационное аннотирование текста подразумевает моделирование структуры выстроенной в нём аргументации (системы связанных аргументов). Аргументы соответствуют утверждениям, сформулированным на языке пропозиций и объединённым отношениями, которые выражают поддержку или опровержение одними утверждениями других (соответственно аргументативными и конфликтными связями).

Связанные утверждения (тезисы) в составе аргумента различаются по ролям: среди них выделяется доказываемое утверждение (заключение), обосновываемое иными (посылками). Роли определяются отдельно для каждого аргумента и могут меняться при связи нескольких аргументов: заключение одного способно служить посылкой в другом. На уровне полного текста определяется главный тезис, выражающий его основную идею.

Разметка аргументационной структуры текста предполагает уточнение семантики связей через их соотнесение со схемами аргументов (для отношений поддержки) или схемами конфликтов (при опровержениях). Схемы соответствуют типовым моделям рассуждений из фиксированного набора. В данной работе применяются схемы из компендиума Уолтона [Walton et al. 2008].

Структура аргументации полного текста представляется в соответствии с форматом AIF (Argument Interchange Format) [Rahwan, Reed 2009] посредством ориентированного графа с двумя типами вершин (информационными для утверждений и вершинами-схемами для схем аргументов и конфликтов). Так, решаемая задача состоит в представлении неструктурированного текста в виде связного аргументационного графа с помощью заданного набора схем рассуждений (связность графа означает наличие не менее одного пути между любыми двумя вершинами).

Аннотирование произведено с помощью набора веб-инструментов, разрабатываемых в ИСИ СО РАН [Сидорова и др. 2020] для создания тематических корпусов текстов, визуализации аргументативных утверждений и схем аргументации. Инструмент предоставляет возможность сохранения графа в формате .json, что позволяет проводить компьютерную обработку аннотаций.

**3. Проблемы и методы аннотирования аргументации**

Моделирование аргументационной структуры каждого текста из корпуса производилось вручную одним разметчиком. Разметка текстов означала решение трёх подзадач, типичных для аннотирования аргументации ([Lawrence, Reed 2019]):

1) *Выявление аргументативных утверждений* через анализ их пропозиционального содержания. При этом в научных и научно-популярных текстах часть аргументов может выражаться таблицами или рисунками, как с интерпретацией на естественном языке (доступной для аннотирования), так и без нее. Неявные тезисы (не представленные эксплицитно) не выделялись при разметке текстов, предназначенных для машинного обучения.

2) *Обнаружение связанных аргументативных утверждений*. Построение связей между тезисами проводилось итеративно от главного тезиса по обновляемому списку не присоединённых к графу утверждений: определялись напрямую атакующие или поддерживающие посылки, для которых затем выявлялись связанные тезисы в постепенно расширяемой анализируемой окрестности текста с учётом уже построенной аргументационной структуры (для выявления связей между позиционно удалёнными утверждениями, представляющего особую сложность).

3) *Уточнение характера построенных связей* за счёт знаний о типовых схемах рассуждений, или детализация графа на уровне вершин-схем. Перспектива компьютерной обработки аннотаций и их использования в машинном обучении влечёт унификацию разметки: ограничение числа используемых схем, обозначение точных критериев выбора одной схемы среди нескольких схожих.

**4. Компьютерная обработка разметки в эксперименте**

Аннотированная часть корпуса включает 110 текстов, из них 10 научных были размечены с целью дальнейшего использования в машинном обучении (5 лингвистических (*Ling*) из области лексикологии с анализом отдельных тематических групп слов, в том числе в культурологическом аспекте, и 5 по компьютерным технологиям (*Comp*) с представлением разработанных решений прикладных задач из областей анализа изображений, интеграции автоматизированных систем, разработки игровых алгоритмов). Другая часть корпуса – тексты научно-популярного жанра (*PS*), аннотированные для изучения приёмов аргументации. Научные статьи взяты из корпуса Ru-RSTreebank [Pisarevskaya et al. 2017] (сведения об их риторических структурах не учитывались). Визуализация подграфа аргументации с иллюстрацией случаев реализации наиболее частотных схем приводится на рис. 1.

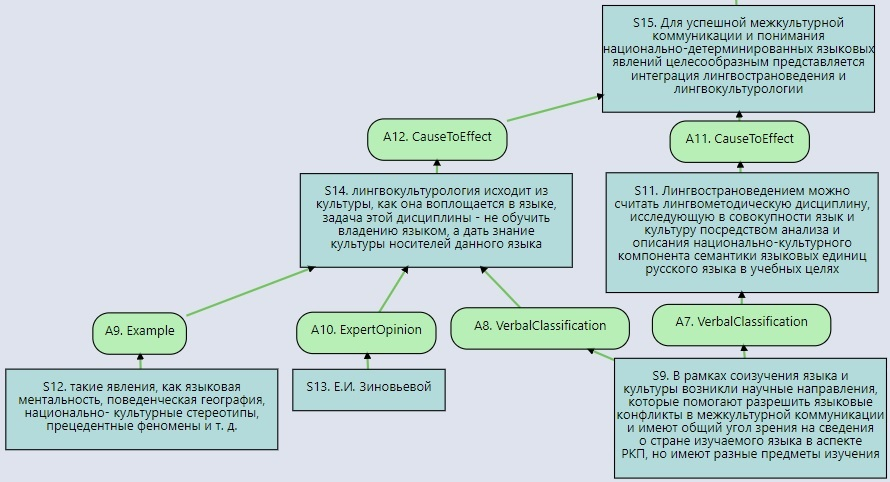


Рис. 1. Пример фрагмента аргументационной разметки.

Сравнительные характеристики каждой из групп текстов корпуса (размер в словах *NW*, число выделенных тезисов *NS*, число построенных аргументативных связей *NA*, количество и доля слов в тезисах *NT*) указаны в табл. 1. Так, при меньшем объёме научных текстов в них выделено большее число тезисов.

*Таблица 1*. Количественные характеристики текстов корпуса.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Comp* | *Ling* | *PS* |
| *NW* | 1149 | 952 | 1535 |
| *NS* | 44 | 45 | 20 |
| *NA* | 45 | 45 | 14 |
| *NT* | 662 (57,6%) | 657 (69,0%) | 316 (20,6%) |

В табл. 2 для каждой группы текстов указываются схемы аргументации с относительной частотой *F* ⩾ *p* (*F*= \*100/*NA*,  – число реализаций i-й схемы в группе, *p* = 5%). Полужирным шрифтом выделены частоты трёх схем с наибольшими *F* в каждой группе текстов, курсивом – схем с *F* < *p* в данной группе.

*Таблица 2*. Встречаемость частотных аргументативных схем.

|  |  |  |  |
| --- | --- | --- | --- |
| Схема | *FComp* | *FLing* | *FPS* |
| Cause to Effect (*от причины к следствию*) | **15%** | 13% | **18%** |
| Correlation to Cause (*через взаимосвязь*) | 14% | **23%** | 5% |
| Example (*от примера*) | **20%** | **27%** | **24%** |
| Expert Opinion (*от мнения эксперта*) | *1%* | 7% | 11% |
| Modus Ponens *(через «правило вывода»*) | - | - | **17%** |
| Negative Consequences (*от осложнений*) | 6% | - | *1%* |
| Positive Consequences (*от выгоды*) | 5% | - | *1%* |
| Practical Reasoning *(от практической цели)* | 12% | 7% | *2%* |
| Sign (*от знака к означаемому*) | *2%* | - | 6% |
| Verbal Classification (*через классификацию*) | **22%** | **20%** | *1%* |
| Суммарные доли приведённых схем: | 98% | 97% | 86% |

Среди аргументационных схем, наиболее часто реализуемых в текстах корпуса, можно выделить группы семантически схожих: обобщённые причинно-следственные доказательства (Cause to Effect, Correlation to Cause), рассуждения практического плана (Practical Reasoning, Positive/Negative Consequences), обоснования через свидетельство (Example, Expert Opinion, Sign).

Сравнение же размеченных научных статей по лингвистике и компьютерным технологиям позволило выявить тематическую зависимость в применении схем. Во-первых, в лингвистических текстах корпуса намного чаще реализуются рассуждения через отсылку к экспертному мнению (схема Expert Opinion) или через глубинный анализ изучаемых сущностей по их внешним проявлениям (Correlation to Cause). Эти отличия объясняются тем, что в представленных текстах по компьютерным технологиям описываются разработки авторов, отчего снижается потребность в цитировании иных исследователей, а созданные программы удобно представлять от их устройства к функционированию.

Во-вторых, прикладной характер статей по компьютерным технологиям часто влечёт анализ практических последствий (и положительных, и отрицательных) от возможных подходов к разработке программы (схемы Positive/Negative Consequences). Рассуждения от практической цели (Practical Reasoning) тоже более свойственны статьям по компьютерным технологиям.

Схожесть двух групп научных текстов проявляется в частом обращении к примерам (схема Example), активном использовании классификаций для анализа различных аспектов исследуемых явлений (Verbal Classification) и причинно-следственных связей общего характера (схема Cause to Effect).

Сравнение частот аргументативных схем позволяет также выявить особенности организации рассуждений в научно-популярных текстах. Во-первых, в их разметке практически не встречается схема Verbal Classification (не требуется высокий уровень систематизации сведений). Во-вторых, доказательства в научно-популярных текстах чаще обращены к авторитетному источнику (ввиду возможности основания работ этого жанра на сопоставлении экспертных воззрений). В-третьих, научно-популярным статьям более свойственно построение общих причинно-следственных связей без глубокой дифференциации (при разметке таких отношений применялась схема Modus Ponens, понимаемая более свободно) ввиду большей обобщённости.

С другой стороны, научные и научно-популярные тексты схожи равной долей использования примеров и реализацией небольшого числа отдельных схем аргументации, что способно помочь отличать их от текстов иных жанров.

**5. Заключение**

Описанный эксперимент позволил выявить особенности аргументационной разметки текстов разных жанров и тематик в рамках академической коммуникации, выделить узкие места в реализации схем аргументации и выработать правила построения аргументационных структур для машинного обучения.

**Литература**

* 1. *Сидорова Е.А., Ахмадеева И.Р., Загорулько Ю.А., Серый А.С., Шестаков В.К.* (2020), Платформа для исследования аргументации в научно-популярном дискурсе. Онтология проектирования, 2020, Т. 10, №4(38), С. 489–502.
  2. *Daxenberger J., Eger S., Habernal I., Stab C., Gurevych I.* (2017), What is the Essence of a Claim? Cross-Domain Claim Identification. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2055–2066.
  3. *Lauscher A., Glavaš G., Ponzetto S.* (2018), An Argument-Annotated Corpus of Scientific Publications. Proceedings of the 5th Workshop on Argument Mining, pp. 40–46.
  4. *Lawrence J., Reed C.* (2019), Argument Mining: A Survey. Computational Linguistics, Vol. 45 (4), pp. 765–818.
  5. *Pisarevskaya D. et al.* (2017),Towards building a discourse-annotated corpus of Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2017», pp. 194–204.
  6. *Rahwan I., Reed C.* (2009), The argument interchange format. Argumentation in artificial intelligence, pp. 383–402.
  7. *Walton D., Reed C., Macagno F.* (2008), Argumentation schemes. New York: Cambridge University Press, 443 p.

**References**

* 1. *Sidorova E.A, Akhmadeeva I.R., Zagorulko Y.A., Seryj A.S., Shestakov V.K.* (2020), Platforma dlja issledovanija argumentacii v nachno-popularnom discurse [Research Platform for the Study of Argumentation in Popular Science Discourse]. Ontologija proektirovaniya [Ontology of Designing], 2020, Vol. 10 (4), pp. 489–502.
  2. *Daxenberger J., Eger S., Habernal I., Stab C., Gurevych I.* (2017), What is the Essence of a Claim? Cross-Domain Claim Identification. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2055–2066.
  3. *Lauscher A., Glavaš G., Ponzetto S.* (2018), An Argument-Annotated Corpus of Scientific Publications. Proceedings of the 5th Workshop on Argument Mining, pp. 40–46.
  4. *Lawrence J., Reed C.* (2019), Argument Mining: A Survey. Computational Linguistics, Vol. 45 (4), pp. 765–818.
  5. *Pisarevskaya D. et al.* (2017),Towards building a discourse-annotated corpus of Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2017», pp. 194–204.
  6. *Rahwan I., Reed C.* (2009), The argument interchange format. Argumentation in artificial intelligence, pp. 383–402.
  7. *Walton D., Reed C., Macagno F.* (2008), Argumentation schemes. New York: Cambridge University Press, 443 p.

**On Argumentation Annotation of Scientific and Popular Science Texts**

**I.S. Pimenov**

**Abstract.** *The task of manual argumentation annotation is approached for scientific and popular science texts in Russian language. The method of building argumentative graphs is implemented with the set of web-tools for creating thematic corpora and visualizing argumentative statements and schemes. Annotation of texts is performed through the use of argumentative schemes from Walton's compendium in three distinct steps (by identifying argumentative statements, analyzing relations between them, specifying exact schemes). Specific difficulties in applying the method to texts of the chosen genres are noted. A quantitative comparison of argumentation structures is performed between scientific texts of different thematic areas and genres.*

**Keywords.** Argumentation annotation, Walton’s argumentation schemes, argumentation graph, scientific and popular science texts.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Пименов Иван Сергеевич**

Новосибирский государственный университет (Россия).

**Pimenov Ivan Sergeevich**

Novosibirsk State University (Russia).

***E-mail: pimenov.1330@yandex.ru***