ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 81'322.2 *Научная статья* DOI: 10.18287/2223-9537-2024-14-1-82-93

Применение методов машинного обучения для выявления аргументативных связей в текстах научной коммуникации

© 2024, Н.В. Саломатина ✉, Е.А. Сидорова, И.С. Пименов

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия

Аннотация

Представлены результаты экспериментов по оценке применимости методов машинного обучения для решения задачи распознавания аргументативных связей в текстах научной коммуникации. Под аргументативной связью понимается отношение, связывающее посылку и заключение типового рассуждения или аргумента, используемого автором для убеждения аудитории. Для оценки качества применялись характеристики точности, полноты и *F*-меры, полученные при решении задачи распознавания аргументативных связей между смежными текстовыми фрагментами двух видов: предложений и клауз. Базой эксперимента послужил русскоязычный корпус текстов из области научной коммуникации с размеченной экспертами-лингвистами аргументацией. Для разметки использован инструмент *ArgNetBank Studio*, позволяющий создавать коллекции текстов с детализированной разметкой аргументации. На основе размеченных текстов построены наборы данных для машинного обучения, в которых соотношение связанных и несвязанных аргументативными отношениями пар фрагментов текста (предложений или клауз) составило 1 к 3. Для повышения качества обучения моделей наборы были сбалансированы двумя способами. В первом случае баланс достигался за счёт того, что из каждого текста отбиралось равное количество пар обоих типов, во втором – пары дублировались. На полученных наборах данных проведены эксперименты по связыванию фрагментов текста методами машинного обучения разных типов. Экспериментально определён диапазон изменения оценок качества при распознавании связанных фрагментов в зависимости от их доли в обучающей и тестовой коллекциях. Установлено, что в рамках существующего дисбаланса в реальных коллекциях значения оценок качества могут изменяться в пределах 40–50%. Новизна работы заключается в исследовании диапазона возможных расхождений в оценках качества при применении разных методов машинного обучения на сбалансированных и несбалансированных обучающих и тестовых коллекциях на русскоязычном материале.

Ключевые слова: *научная коммуникация, анализ аргументации, аргументативная разметка текста, аргументативные отношения, методы машинного обучения.*

Цитирование: *Саломатина Н.В., Сидорова Е.А., Пименов И.С.* Применение методов машинного обучения для выявления аргументативных связей в текстах научной коммуникации // Онтология проектирования. 2024. Т.14, №1(51). С.82-93. DOI: 10.18287/2223-9537-2024-14-1-82-93.

Финансирование: исследование выполнено за счёт гранта Российского научного фонда № 23-11-00261, https://rscf.ru/project/23-11-00261/.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Во многих приложениях, осуществляющих автоматический анализ текстов, важно учитывать аргументационную составляющую, в частности, для оценки их убедительности, понимания и ведения дискуссий, принятия решений в рекомендательных системах и т.д. В этих случаях структура аргументации должна распознаваться автоматически. Для решения этой

задачи преимущественно применяются методы машинного обучения (МО), что предполагает наличие корпусов с аргументационной разметкой. Разметка аргументации трудоёмка, требует экспертных навыков семантического и прагматического анализа текстов и, как следствие, не может быть осуществлена с помощью краудсорсинга. Существующие наборы данных имеют небольшой объём, а наличие нескольких моделей аргументации, употребляемых для разметки разными группами исследователей, осложняет решение задачи. Поэтому, несмотря на преимущества глубокого обучения, часто (особенно в случае дефицита данных) используют традиционные методы МО, которые в ряде случаев показывают сравнимые результаты с полученными при применении нейронных сетей (НС).

Автоматическое извлечение отдельных аргументов и построение аргументативной структуры текста выполняется в несколько этапов, которые также проходит аннотатор при ручной разметке аргументации [1], в частности: 1) выявление фрагментов текста, содержащих аргументацию; 2) построение связей между аргументативными фрагментами (распознавание аргументов); 3) уточнение ролей фрагментов в составе аргумента (посылок, выводов); 4) выявление типов аргументов (схем аргументации).

Установление связей между фрагментами является самой сложной задачей в области анализа аргументации [1, 2]. Наилучшие результаты достигаются при существенных ограничениях, таких как ограничение определённой предметной областью, заданной сегментацией, поиском связей с заданным тезисом и т.п. Без вводимых ограничений качество получаемых результатов, как правило, значительно ниже. Важную роль в систематизации различных исследований играет наличие эталонного или базового решения (ЭР), т.е. решения, полученного традиционными методами без привлечения какой-либо дополнительной информации. Применительно к задаче извлечения аргументативных отношений это означает отсутствие какой-либо информации о фрагментах текста, кроме наличия/отсутствия связи между ними. Такие решения могут в дальнейшем служить оценкой эффективности других разрабатываемых методов, в которых часто находят применение различные дополнительные признаки.

Одной из особенностей, делающих данную задачу труднорешаемой, является отсутствие формальных критериев выделения границ утверждений, входящих в аргументацию. Обычно в исследованиях рассматривают фрагменты-предложения. Однако сложные предложения могут содержать аргументы внутри себя, и если рассматривать в качестве фрагмента целое предложение, то аргумент, содержащийся внутри, будет утрачен. Использование клауз (конструкций с единственным предикативным элементом) позволяет рассматривать более подробную аргументативную разметку, но делает создание размеченных корпусов более трудоёмким, а наборы данных менее сбалансированными.

В настоящей работе проводится сравнение методов решения этой задачи на материале размеченного корпуса русскоязычных текстов из области научной коммуникации. Цель исследования – получить оценки качества для традиционных методов МО на этапе связывания утверждений в аргумент посредством обработки фрагментов двух типов (предложений и клауз), установить диапазон изменения оценок качества для сбалансированных и несбалансированных коллекций. Использовались три алгоритма классификации, часто применяемые в исследованиях по извлечению аргументации из текстов: полиномиальный наивный Байес (*MNB*), метод опорных векторов (*SVM*), многослойный перцептрон (*MLP*).

1 Подходы к распознаванию аргументативных связей

В работах по распознаванию в тексте аргументации, в т.ч. и аргументативных связей, ЭР не указывается вовсе или вычисляется авторами с учётом особенностей каждого эксперимента, что затрудняет сравнение полученных результатов с другими исследованиями. Чаще

представляются подходы, опирающиеся на использование дополнительной информации: о модели (схеме) рассуждения [3, 4], о структуре (роли компонентов) аргумента [5], о дискурсивных маркерах и аргументативности фрагментов [6], о маркерах дискурса и тематической структуре текста [7, 8], о контекстной информации и теме [9].

В ряде работ проводится сравнение методов на основе глубокого обучения и традиционных методов МО. Так, в работе [2] приведены оценки по *F*-мере, достигнутые по результатам обнаружения аргументативных связей в *CDCP*-корпусе[1] (более 700 комментариев пользователей о практике взыскания долгов) с помощью рекуррентных НС. В одном случае НС уступает по эффективности обнаружения аргументативных связей алгоритму *SVM* с векторными представлениями *GloVE*[2], в другом превосходит его [10]. При этом ЭР в явном виде не фиксируются, достаточным считается сравнение алгоритмов на одном и том же корпусе.

Результаты применения модели *ruRoberta*[3] и полносвязных НС для связывания смежных фрагментов из русскоязычных научных и научно-популярных текстов в аргумент с известной информацией об аргументативности хотя бы одного фрагмента и наличии маркера дискурса в тексте представлены в работе [6].

Поиск аргументативных связей может опираться на распознавание модели рассуждения (схемы). Обнаружение такой модели означает, что связь между составляющими этого рассуждения установлена. На базе разнотипных лексических констант (маркеров дискурса и др.) строятся лексико-синтаксические шаблоны, поиск по которым позволяет решать задачу извлечения аргументов комплексно: определять связанные фрагменты, устанавливать тип связи и роли фрагментов (компонентов аргумента). Исследование, проведённое в работе [3] на материале английского языка, посвящено автоматической классификации аргументов по пяти схемам, наиболее частотным в обрабатываемой коллекции: «от примера» (*Example*), «причина-следствие» (*Cause_to_Effect*), «от практической цели» (*Practical Reasoning*) и др. Набор признаков включал как общие для всех схем признаки (позиционные характеристики, длину интервала между ними в тексте и пр.), так и особенности каждой схемы (от ключевых слов и знаков препинания до синтаксических зависимостей). Применён алгоритм дерева решений, оценки качества работы которого зависели от анализируемой схемы: значения аккуратности составляли от 60 до 90%, тогда как значения ЭР, достигнутые на общих признаках, равны 50%.

Для поиска в русскоязычных текстах рассуждения «от экспертного мнения» в [4] разработан шаблон, точность распознавания по которому составила 86,5%. Известны шаблоны, работающие с точностью 75%, 91% и 86% для поиска рассуждений «от примера», «по аналогии», «согласно классификации», соответственно, в текстах на русском языке [11]. Использование в этих шаблонах дискурсивных маркеров потребовало проведения предобработки текста, а именно, классификации фрагментов на аргументативные и неаргументативные. Значения ЭР не указывались, сравнение результатов с ними не проводилось.

Недостатком комплексного решения задачи поиска связей с использованием шаблонов является трудоёмкость их создания, а поиск по ним не гарантирует построения полного связного графа рассуждений в тексте.

Знания об аргументативности и роли фрагментов использовались при установлении связи в работе [5]. Клаузы, предварительно выделенные в тексте эссе, классифицировались по их типу: главное утверждение, утверждение, посылка, неаргументативное утверждение. Свя-

[1] *Cornell eRulemaking Corpus – CDCP* - это корпус для анализа аргументов, снабжённый информацией о структуре аргументации, отражающей возможность оценки аргументов. https://paperswithcode.com/dataset/cdcp.

[2] *Global Vectors for Word Representation - GloVe* - это неконтролируемый алгоритм обучения для получения векторных представлений слов. https://nlp.stanford.edu/projects/glove/.

[3] *ruRoBERTa large* - русская языковая модель, которая может определять вероятности следующего и пропущенного слова и эффективно представлять слова и тексты в векторном пространстве. https://cloud.ru/ru/datahub/rugpt3family/ruroberta-large.

зывание клауз осуществлялось посредством бинарной классификации: все возможные в рамках одного абзаца пары идентифицировались как связанные или нет. Лучшие результаты достигнуты с помощью метода *SVM* с *F*-мерой равной 72%. Эффективными признаками признаны лексические (пары слов, первое слово в утверждении и модальные слова), синтаксические (продукционные правила, извлечённые из дерева синтаксического разбора) и маркеры дискурса. Значения ЭР определены исходя из критерия Мак-Немара.

Результаты экспериментов, представленные в [9], показывают, что признаки из экспериментов в работе [5], обогащённые тематическими словами, общими словами из контекста аргументативных утверждений (рассматриваются предложения внутри абзаца), дискурсивными маркерами эффективны для установления связей и их типов на уровне наличия связи или её отсутствия, типа связи («поддержки» или «атаки»). Эксперименты на данных студенческих эссе демонстрируют *F*-меру 75% в первом случае и 67% – во втором. Значениями ЭР служат оценки качества, полученные в [5].

В статье [8] связанность аргументативных утверждений определяется с помощью автоматически генерируемых тематических моделей. Модели формируются на основе предложений с маркерами, позволяющими определить роли фрагментов предложения, из документов, найденных в сети Интернет по ключевым словам, извлечённым из предварительно распознанных как аргументативные утверждения текста. На основе пар посылка–заключение генерируется тематическая модель, на основе которой строится матрица вероятностей соотношения темы посылки и темы заключения. По полученной матрице для каждой пары утверждений исходного текста рассчитывается вероятность того, что эти утверждения связаны. Получены следующие оценки качества: по точности – 60%, полноте – 82%, *F*-мере – 69%. В качестве базовых значений приняты значения, соответствующие случайным – 50%.

Большинство исследований по анализу аргументации выполнено на текстах английского языка. Обзор работ показывает, что общепринятого подхода к вычислению ЭР нет. В условиях дефицита русскоязычных коллекций с детализированной разметкой аргументации в данной статье для формирования ЭР использованы традиционные методы МО, показавшие наилучшие решения в рассмотренных работах, и простейшее представление вектора данных без использования дополнительных сведений о свойствах этих данных.

2 Моделирование аргументации

Обучающая и тестовая коллекции, исследуемые в эксперименте, содержат тексты с экспертной разметкой (аннотацией) аргументации. Аннотирование каждого текста заключается в построении формального представления его аргументативной структуры, которая объединяет все приводимые в тексте аргументы. В текстовом оформлении аргумент выражается набором связанных утверждений (фрагментов текста на естественном языке), где все утверждения (называемые посылками), кроме одного (заключения), обосновывают это одно утверждение (либо, в случае атаки на это утверждение, опровергают его). Связь утверждений внутри аргумента соответствует реализации конкретной модели рассуждения. Посылка и заключение одного аргумента могут быть посылкой или заключением другого. Связи между утверждениями позволяют объединить их в аргументативную структуру текста.

В качестве примера аргументативной разметки рассмотрен абзац, взятый из рецензии на научную статью (предложения пронумерованы для наглядности):

(1) *Непоследовательность прослеживается и в выводах.* (2) *Первый («В результате анализа было выявлено, что практически все игровые виды спорта пользуются популярностью в качестве рекреационного времяпрепровождения».) в работе вообще не исследовался.* (3) *А второй («Они отражают географические названия тех местностей и городов, где они впервые появились и доносят разного рода информацию».) носит предельно «школьный» характер и не заслуживает доведения до внимания научного сообщества.*

Данный абзац содержит три предложения, где:

- (1)-(2) между первым (заключением) и вторым (посылкой) предложениями выстроена связь по аргументативной схеме часть–целое;
- (2)-(3) между вторым и третьим предложениями нет непосредственной аргументативной связи;
- (1)-(3) третье предложение содержит посылку к первому по аналогичной аргументативной схеме (не является посылкой либо заключением ко второму предложению).

При обработке на уровне клауз в абзаце выделяются восемь фрагментов (одна клауза в первом предложении, три во втором, четыре в третьем), где при аргументативной разметке указана дополнительная связь между двумя последними фрагментами:
от: *А второй носит предельно «школьный» характер* (посылка)
к: *и не заслуживает доведения до внимания научного сообщества* (заключение)
аргументативная схема: «апелляция к личности».

Моделирование аргументации соответствует стандарту формата обмена аргументами (*Argument Interchange Format, AIF*) [12]. *AIF* определяет основные этапы аргументативной разметки: идентификация утверждений, обнаружение связей между ними, указание модели рассуждения для каждой связи. Выявление связей включает уточнение роли утверждений в аргументах (какие утверждения являются посылками, а какие – заключением). Специфика моделей рассуждений предполагает их выбор из принятой классификации, в частности из сборника схем аргументации Уолтона [13], применимого к текстам различных жанров и рекомендованного разработчиками *AIF*.

Аргументативная разметка текстов, используемая в эксперименте, выполнена экспертами с использованием онлайн-платформы *ArgNetBank Studio* (*https://uniserv.iis.nsk.su/arg*) [14].

3 Выявление аргументативных связей

Решается задача классификации множества смежных фрагментов:

$FR = \{fr_i \,\|\, fr_{i+1}\}$, где i – позиция фрагмента в тексте,

$FR = FR^L \cup FR^T$, где FR^L – обучающая коллекция, FR^T – тестовая коллекция,

на два класса из множества категорий:

$R = \{r^+, r^-\}$, где r^+ – класс связанных, а r^- – несвязанных фрагментов.

Для смежных фрагментов в обучающей коллекции $FR^L \subset FR$ известны метки из R. На ней строится классификатор F: $FR \times R \rightarrow$ {истина, ложь}.

Для получения базовых оценок качества распознавания связи в данном исследовании выбирается простейшее векторное представление фрагментов леммами. Для уменьшения размерности проводится автоматическая фильтрация лемм по частоте и формальному критерию χ^2, позволяющему устранить из вектора леммы, малоинформативные для установления связности фрагментов. Пороги по информативности признаков определяются экспериментально. Для исследования выбраны три алгоритма классификации: *MNB*, *SVM* и *MLP*. Использовались программные реализации алгоритмов на *Python* из библиотеки [15].

3.1 Экспериментальное исследование выявления аргументативных связей

Оценки качества получены на корпусе из 146 русскоязычных текстов, относящихся к области «научная коммуникация», а именно: рецензии на научные статьи, короткие научные статьи по информационным технологиям и лингвистике, новости науки, аналитические статьи с сайта Хабр [16]. К разметке корпуса были привлечены четыре аннотатора - специали-

сты в области лингвистики, в т.ч. компьютерной. Аннотированный корпус содержит 10295 предложений или 27159 клауз.

Оценка согласия между аннотаторами проведена на подкорпусе из 50 текстов с дублированной разметкой по алгоритму из работы [17], не учитывающему случайные совпадения в разметке. Сравнение аннотаций, построенных разными экспертами для одних и тех же текстов, показывает, что доля совпадающих связей, определённых по совокупности совпадающих утверждений (их доля равна 83%), достигает 55%. Это соответствует нижней границе, поскольку отдельные формально несовпадающие связи в разных конфигурациях, параллельных или последовательных, являются допустимыми различиями.

Корпус, на котором проведены эксперименты, содержал одну версию аннотации для каждого текста. Корпус разделялся на обучающую и тестовую коллекции в пропорции 80% и 20% соответственно. Данные для коллекций формировались путём прохода по тексту скользящим окном шириной в два фрагмента (предложения или клаузы). Распределение фрагментов по коллекциям сохраняло целостность текстов (любые два фрагмента из одного и того же текста принадлежат одной и той же коллекции), поскольку эксперименты показали, что нарушение этого принципа существенно завышает показатели качества.

Соотношение числа смежных и несмежных связей в коллекции зависит от многих факторов (жанра и темы текста, стиля изложения аргументации автором, стиля разметчика), влияющих на выбор схем при построении аргументации, которые определяют контактность/ неконтактность посылок и заключения. Разбалансировка связанных и несвязанных пар в построенной коллекции достигала соотношения 1 к 3 в пользу несвязанных. Чтобы получить возможный диапазон оценок качества, проведены эксперименты как с несбалансированными вариантами тестовых и обучающих коллекций (FR^{L}), так и со сбалансированными: в коллекции FR^{L1} оставлено оригинальное соотношение связанных и несвязанных пар; в FR^{L2} и FR^{L3} соотношение пар сбалансировано. Баланс достигнут за счёт того, что: в первом случае из каждого текста отбиралось равное количество пар обоих типов (если число пар одного типа превышало число для другого, то из текста отсеивались случайные пары первого типа сверх их количества для второго); во втором – дублировались пары связанных фрагментов.

Для создания векторных представлений пар фрагментов в качестве признаков использованы леммы. Значения компонентов вектора – бинарные (соответствуют встречаемости/ отсутствию леммы во фрагменте). Векторизация (на этапах обучения и распознавания) заключалась в построении отдельных векторов для каждого из двух фрагментов и их последующей конкатенации в общий вектор. Раздельная обработка обоих фрагментов позволила учитывать позиционную специфику лемм, их влияние на наличие/отсутствие связи с соседним предложением в зависимости от контекста (располагается соседнее предложение справа или слева). Позиционная специфика лемм учитывалась и при формальной фильтрации признаков: наборы признаков строились отдельно для обеих позиций (слева/справа) с учётом: во-первых, частоты лемм (встречаемости во фрагментах соответствующей позиции, как минимум в 5 предложениях либо в 10 клаузах для каждого из двух типов фрагментов); во-вторых, их связи с распределением классов (по критерию χ^2: для обоих типов фрагментов выбирались 20% лемм, наиболее информативных в указании на класс пары).

Результаты эксперимента с одинаковым распределением по двум классам (наличие или отсутствие аргументативной связи) приведены в таблице 1. Полужирным шрифтом выделены лучшие значения оценок качества. Использованы обозначения: *P*, *R* – точность и полнота соответственно. По точности идентификации аргументативной связи лучшие результаты показал метод *SVM*, по полноте – *MNB*. Оценки качества, полученные для предложений и клауз, в целом близки (максимум расхождения: 6% по точности, 15% по полноте, 7% по *F*-мере), для предложений они часто выше, в т.ч. для сбалансированных коллекций. Вид балан-

сировки обучающей коллекции сказался, в основном, на показателях полноты и, как следствие, *F*-меры. Балансировка коллекций дублированием недостающих экземпляров класса менее эффективна, чем фильтрация лишних элементов, проведённая случайным образом.

Таблица 1 – Результаты экспериментов с одинаковым распределением в обучающей и тестовой коллекции

		Точность (*P*)			**Полнота (*R*)**			***F*-мера**		
		MNB	*SVM*	*MLP*	*MNB*	*SVM*	*MLP*	*MNB*	*SVM*	*MLP*
FR^{L1}	предл.	0.36	**0.45**	0.31	0.15	0.13	**0.27**	0.21	0.20	**0.29**
	клаузы	0.34	**0.39**	0.29	**0.12**	0.09	**0.12**	0.17	0.14	**0.27**
FR^{L2}	предл.	0.57	**0.59**	0.57	**0.57**	**0.57**	0.55	0.57	**0.58**	0.56
	клаузы	0.55	**0.63**	0.56	**0.59**	0.58	0.54	0.57	**0.61**	0.55
FR^{L3}	предл.	0.48	**0.64**	0.56	**0.45**	0.34	0.27	**0.46**	0.45	0.37
	клаузы	0.53	**0.61**	0.57	**0.40**	0.27	0.29	**0.45**	0.38	0.39

Для случая неодинакового распределения связанных и несвязанных пар фрагментов в обучающей и тестовой коллекциях проведён эксперимент, позволяющий получить представление о нижней границе оценок качества: обучение проведено на сбалансированных коллекциях, в тестовой коллекции сохранена свойственная всей коллекции диспропорция (см. таблицу 2).

Сравнение оценки качества в таблице 1 варианта FR^{L1} с результатами в таблице 2 показало, что разница в распределениях пар фрагментов в обучающей и тестовой коллекциях влечёт за собой некоторое понижение точности распознавания связей, но значительно повышает полноту при обучении на сбалансированной коллекции.

Таблица 2 – Результаты экспериментов с неодинаковым распределением в обучающей и тестовой коллекции

		Точность (*P*)			**Полнота (*R*)**			***F*-мера**		
		MNB	*SVM*	*MLP*	*MNB*	*SVM*	*MLP*	*MNB*	*SVM*	*MLP*
FR^{L2}	предл.	0.28	**0.33**	0.29	**0.57**	0.56	0.56	0.37	**0.42**	0.38
	клаузы	0.25	**0.31**	0.25	**0.59**	**0.59**	0.55	0.35	**0.41**	0.35
FR^{L3}	предл.	0.24	**0.38**	0.30	**0.45**	0.34	0.27	0.31	**0.36**	0.29
	клаузы	0.27	**0.34**	0.31	**0.40**	0.27	0.29	**0.32**	0.30	0.30

При смещении баланса в коллекции в сторону связанных пар оценки качества их распознавания могут возрасти, но в реальных текстах научной коммуникации случаи с дисбалансом в пользу связанных смежных пар маловероятны. По данным таблиц можно установить примерный диапазон изменения показателей качества:

- для предложений: $0.24 \le P \le 0.64$, $0.13 \le R \le 0.57$, $0.20 \le F\text{-мера} \le 0.58$;
- для клауз: $0.25 \le P \le 0.63$, $0.09 \le R \le 0.59$, $0.14 \le F\text{-мера} \le 0.57$.

3.2 Анализ результатов

Анализ результатов распознавания связей позволил выявить три основных типа ошибок, общих для обоих видов фрагментов.

3.2.1 Ошибки, связанные с узким контекстом рассматриваемых связей

Наиболее распространённый случай ошибочного распознавания связей обусловлен спецификой аргументативного аннотирования целостных текстов. При ручной разметке аргументации в тексте эксперты учитывают широкий контекст даже для выделения отдельных связей: ориентируются на логическую организацию текста, анализируют роль частных утверждений и связей между ними в доказательстве автором его ключевых тезисов. Автоматическое выявление связей в эксперименте ограничено предельно узким контекстом: непо-

средственно парой фрагментов, проверяемых на наличие объединяющей их аргументативной связи. Для данного типа ошибок можно выделить два основных подтипа в зависимости от специфики аргументативного аннотирования фрагментов в исходном тексте.

1. Классификаторы (зачастую все три алгоритма сразу) устанавливают связь между фрагментами ввиду наличия в них явных дискурсивных маркеров такой связи, однако аннотатор встроил эти фрагменты в общую аргументативную структуру полного текста иным образом, без построения связи между этой парой фрагментов (например, ввиду их присоединения как параллельных посылок к некоторому третьему фрагменту). Пример подобного расхождения:

(1) *7zip – популярный архиватор с открытым исходным кодом, который получил широкое распространение благодаря своей высокой степени сжатия данных и поддержке множества форматов архивов.* (2) *Он стал неотъемлемым инструментом для многих пользователей, которые сталкиваются с необходимостью архивации и извлечения данных.* (3) ***Однако****, как и любая другая программа, 7zip не застрахован от уязвимостей, которые могут стать угрозой для безопасности пользователей.*

Третье предложение абзаца (а также первая клауза этого предложения) начинается с дискурсивного маркера «однако», который при употреблении в начале фрагмента, как правило, указывает на противопоставление этого фрагмента предыдущему и наличие между ними аргументативной связи опровержения по схеме «логический конфликт». Но, поскольку второе предложение содержит посылку к первому (дополняет тезис о широком использовании и преимуществах архиватора через указание цели, для которой архиватор используется), аннотатор при разметке текста построил связь от третьего предложения к первому (пропуская второе как частный случай первого в контексте указания недостатка архиватора независимо от цели его использования).

2. Близким случаем к альтернативному построению связи выступает иной подход аннотатора к детализации текстового фрагмента, когда два фрагмента, между которыми классификаторы выявляют предполагаемую связь, были объединены разметчиком в одно целостное аргументативное утверждение (например, когда оба фрагмента дополняют друг друга в роли одной смысловой посылки к другому утверждению, а при пропуске любого из них другой не смог бы функционировать как самостоятельная посылка из-за смысловой неполноты). Такое объединение нескольких фрагментов (в т.ч. предложений) в одно целостное утверждение без детализации внутренних связей встречается, например, при приведении цитат. Так, следующий фрагмент текста содержит цитату из двух предложений, объединённых вместе для обоснования первого предложения, однако все три классификатора определили эту пару предложений внутри цитаты как содержащую аргументативную связь.

(1) *Также российский учёный по сути сформулировал гипотезу лингвистической относительности.* (2) *Вот что он писал:* (3) *«При посредстве слов мы думаем и о том, что без тех или других знаков не могло бы быть представлено в нашем мышлении, и точно так же при посредстве слов мы получаем возможность думать так, как не могли бы думать при отсутствии знаков для мышления по отношению именно к обобщению и отвлечению предметов мысли.* (4) *Знаки языка для мысли становятся в процессе речи знаками для выражения мысли или её части, именно – непосредственно знаками для выражения мысли или её части, в состав которой входят представления произносимых слов».*

3.2.2 Ошибки, связанные с отсутствием предварительной обработки текста

Другой тип ошибок обусловлен спецификой постановки задачи для проведения экспериментов: оценивается эффективность распознавания аргументативных связей без проведения отдельных этапов предварительной обработки текста. Пропускаемые этапы затрагивают не только разграничение аргументативных и неаргументативных предложений, но и отражение логической организации целостного текста (его членения на абзацы и логические разделы). Без учёта этой структуры фрагменты объединяются в пары сплошным потоком (так, в пару объединяются последнее предложение одного абзаца и начальное предложение следующего,

а также заголовок раздела текста, если он образован самостоятельным предложением, и первое предложение этого раздела).

(1) *Что такое exclude rules в 7zip*

(2) *Exclude rules в 7zip представляют собой набор правил, которые позволяют пользователям исключать определённые файлы или каталоги при архивации или извлечении данных.*

Данный пример содержит пару из заголовка раздела и его первого предложения, которые определены классификаторами как связанные, тогда как аннотатор изначально не выделил заголовок как отдельное аргументативное утверждение и не включил его в аргументативную структуру текста.

3.2.3 Ошибки неправильной сегментации

Третий случай некорректного распознавания аргументативных связей (в частности, пропуска классификаторами связей, установленных аннотатором) обусловлен остаточными ошибками от сегментации текстов (производимой автоматически).

Из анализа ошибок можно заключить, что полезно применять предобработку текста с целью фильтрации заведомо несвязанных пар, а также с целью извлечения дополнительной информации о фрагментах, в частности об их аргументативности, о маркерах дискурса (например, путём регулирования их веса в векторе представления фрагмента) и пр.

Заключение

В данной работе представлены результаты апробации трёх разнотипных методов МО, использованных для распознавания аргументативных связей в текстах русского языка. Эксперименты проведены на материале текстов разных жанров из области научной коммуникации. Для классификации выбирались смежные фрагменты двух типов: предложения и клаузы, что позволило учесть степень подробности аргументативной разметки. Наилучшие результаты показали методы *SVM* (по точности) и *MNB* (по полноте), при этом для предложений и клауз оценки качества, в целом, близки. Построен диапазон возможных расхождений в оценках качества при применении разных методов обучения на сбалансированных/несбалансированных обучающих и тестовых коллекциях. Результаты показали значимость сбалансированности обучающей коллекции для параметра полноты. Анализ ошибок распознавания связей выявил необходимость предобработки текста и учёта более широкого контекста при выборе пар для классификации.

Полученные оценки качества можно рассматривать как ЭР, поскольку в данном исследовании не применялись дополнительные способы повышения качества, такие как предварительная лингвистическая обработка текста, выделение главного тезиса, использование индикаторов. Полученные результаты могут быть полезны для анализа эффективности вновь разрабатываемых методов, в том числе на основе НС подходов.

Список источников

[1] ***Lawrence J., Reed C.*** Argument Mining: A Survey. Computational Linguistics, 2019. Vol. 45 (4). P.765-818.

[2] ***Chen T.*** BERT Argues: How Attention Informs Argument Mining. Honors Theses, vol. 1589. 2021.

[3] ***Feng V.-W., Hirst G.*** Classifying arguments by scheme // In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011. Vol.1. P.987-996.

[4] ***Achmadeeva I., Kononenko I., Salomatina N., Sidorova E.*** Indicator Patterns as Features for Argument Mining // In: Proc. of the Int. Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). Novosibirsk, 2019. P.0886-0891. DOI: 10.1109/SIBIRCON48586.2019.8958295.

[5] ***Stab C., Gurevych I.*** Identifying argumentative discourse structures in persuasive essays // In: Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, 2014. P.46–56.
[6] ***Sidorova E., Akhmadeeva I., Kononenko I., Chagina P.*** The role of Indicators in Argumentative Relation Prediction // In: Proc. of the Int. Conf. on Computational Linguistics and Intellectual Technologies "Dialogue 2023". Issue 22. 2023. P.477-485. DOI:10.28995/2075-7182-2023-22-477-485.
[7] ***Lawrence J., Reed C.*** Combining argument mining techniques // In: Proc. of the 2nd Workshop on Argumentation Mining. Denver: Association for Computational Linguistics, 2015. P.127-136.
[8] ***Lawrence J., Reed C.*** Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models // In: Proc. of the 4th Workshop on Argument Mining. Denmark: Association for Computational Linguistics, 2017. P.39-48.
[9] ***Nguyen H.V., Litman D.*** Context-aware argumentative relation mining // In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin. 2016. Vol. 1: Long Papers. P.1127-1137.
[10] ***Niculae V., Park J., Cardie C.*** Argument mining with structured SVMs and RNNs // In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Vol. 1: Long Papers. P.985-995.
[11] ***Zasypkin A.S., Pimenov I.S., Salomatina N.V***. The Combined Approach to Identifying Argumentation Structures in Short Scientific Papers // In: IEEE 24th International Conference of Young Professionals in Electron Devices and Materials (EDM). 2023. P.1800-1805. DOI:10.1109/EDM58354.2023.10225223.
[12] ***Rahwan I., Reed C.*** The argument interchange format // In: G. Simari, I. Rahwan (ed.): Argumentation in Artificial Intelligence. Boston: Springer, 2009. P.383-402.
[13] ***Walton D., Reed C., Macagno F.*** Argumentation schemes. Cambridge University Press, 2008. 443 p.
[14] ***Сидорова Е.А., Ахмадеева И.Р., Загорулько Ю.А., Серый А.С., Шестаков В.К.*** Платформа для исследования аргументации в научно-популярном дискурсе // Онтология проектирования. 2020. Т.10, №4(38). С.489-502. DOI: 10.18287/2223-9537-2020-10-4-489-502.
[15] Scikit Learn Homepage. https://scikit-learn.org/stable/supervised_learning.html.
[16] Сайт Хабр. https://habr.com/ru/articles/
[17] ***Пименов И.С.*** Анализ расхождений в аргументационной разметке научных статей на русском языке // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2023. Т.21, №2. С.89-104. DOI: 10.25205/1818-7935-2023-21-2-89-104.

Сведения об авторах

Саломатина Наталья Васильевна, 1958 г. рождения. Окончила Новосибирский государственный университет (НГУ) в 1980 г., к.ф.-м.н. (2009). Старший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А.П. Ершова (Новосибирск). В списке научных трудов более 80 работ в области распознавания речи, анализа символьных последовательностей, компьютерной лингвистики. Author ID (РИНЦ): 5683; ORCID: 0000-0001-8412-9116; Author ID (Scopus): 57190173916; Researcher ID (WoS): G-3032-2019. *salomatina_nv@live.ru*. ✉

Сидорова Елена Анатольевна, 1977 г. рождения. Окончила НГУ в 2000 г., к.ф.-м.н. (2006). Старший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А.П. Ершова (Новосибирск), доцент кафедры программирования и кафедры систем информатики НГУ, член Российской и Европейской ассоциаций искусственного интеллекта. В списке научных трудов более 160 работ в области компьютерной лингвистики, онтологического инжиниринга и разработки интеллектуальных систем. Author ID (РИНЦ): 146000; ORCID: 0000-0001-8731-3058; Author ID (Scopus): 41961707000; Researcher ID (WoS): K-2432-2018. *lsidorova@iis.nsk.su*.

Пименов Иван Сергеевич, 1997 г. рождения. Окончил магистратуру НГУ в 2020 г. Программист 2 категории в Институте систем информатики им. А.П. Ершова (Новосибирск), аспирант кафедры фундаментальной и прикладной лингвистики НГУ. В списке научных трудов 15 работ в области компьютерной лингвистики, в том числе автоматического анализа аргументации. Author ID (РИНЦ): 1164941, ORCID: 0000-0001-5946-9469. *pimenov.1330@yandex.ru*.

Поступила в редакцию 29.11.2023, после рецензирования 18.01.2024. Принята к публикации 2.02.2024.

Scientific article DOI: 10.18287/2223-9537-2024-14-1-82-93

Applying machine learning methods to identify argumentative connections in scientific communication texts

© 2024, N.V. Salomatina ✉, E.A. Sidorova, I.S. Pimenov

A.P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, Novosibirsk, Russia

Abstract

The paper presents the results of experiments to assess the machine learning methods applicability for solving the problem of identifying argumentative connections in scientific communication texts. Argumentative connection is understood as a relationship that connects the premise and the conclusion of a typical reasoning or an argument used by the author to persuade the readers. To assess the quality, the characteristics of accuracy, completeness and F-measure were used obtained when solving the problem of recognizing argumentative connections between adjacent text fragments of two types: sentences and clauses. The basis of the experiment was a Russian-language corpus of texts from the field of scientific communication with arguments marked up by linguistic experts. For markup, the ArgNetBank Studio tool was used, which allows creating collections of texts with detailed argumentation markup. Data sets for machine learning were built on the basis of labeled texts, in which the ratio of pairs of text fragments (sentences or clauses) connected and non-connected by argumentative relationships was 1 to 3. To improve the quality of model training, the sets were balanced in two ways. In the first case, a balance was achieved due to the fact that an equal number of pairs of both types were selected from each text; in the second, pairs were duplicated. Using the obtained data sets, experiments were carried out on linking text fragments using different types of machine learning methods. The range of changes in quality assessments when recognizing related fragments depending on their share in the training and test collections was experimentally determined. It has been established that, within the framework of the existing imbalance in real collections, the values of quality assessments can vary within 40–50%. The novelty of the work lies in the study of the range of possible discrepancies in quality assessments when applying different machine learning methods on balanced and unbalanced training and test collections in Russian-language material.

Keywords: *scientific communication, argumentation analysis, argumentative text markup, argumentative relationships, machine learning methods.*

For citation: *Salomatina NV, Sidorova EA, Pimenov IS.* Applying machine learning methods to identify argumentative connections in scientific communication texts [In Russian]. *Ontology of designing.* 2024; 14(1): 82-93. DOI: 10.18287/2223-9537-2024-14-1-82-93.

Financial Support: The study was supported by the Russian Science Foundation grant No. 23-11-00261, https://rscf.ru/project/23-11-00261/.

Conflict of interest: The authors declare no conflict of interest.

List of tables

References

[1] ***Lawrence J, Reed C.*** Argument mining: A survey. Int. J. of Computational Linguistics 2019; 45(4): 765-818.
[2] ***Chen T.*** BERT Argues: How Attention Informs Argument Mining. Honors Theses; 2021; 1589.
[3] ***Feng V-W, Hirst G.*** Classifying arguments by scheme. In: Human Language Technologies, proc. of the 49th Annual Meeting of the Association for Computational Linguistics: 2011; 1: 987-996.
[4] ***Akhmadeeva IR, Kononenko IS, Salomatina NV, Sidorova EA.*** Indicator Patterns as Features for Argument Mining. In: Engineering, Computer and Information Sciences, proc. of the Int. Multi-Conference SIBIRCON (Novosibirsk). 2019: 0886-0891. DOI: 10.1109/SIBIRCON48586.2019.8958295.

[5] ***Stab C, Gurevych I.*** Identifying Argumentative Discourse Structures in Persuasive Essay. Empirical Methods in Natural Language Processing (EMNLP): Proc. of the Int. Conf. (Doha, Qatar); 2014: 46–56.
[6] ***Sidorova EA, Akhmadeeva IR, Kononenko IS, Chagina PM.*** The role of Indicators in Argumentative Relation Prediction. In: Computational Linguistics and Intellectual Technologies, proc. of the Int. Conf. "Dialogue 2023". 2023; 22: 477-485. DOI: 10.28995/2075-7182-2023-22-477-485.
[7] ***Lawrence J, Reed C.*** Combining argument mining techniques. In: Argumentation Mining, proc. of the 2nd Workshop (Denver). Association for Computational Linguistics, 2015: 127-136.
[8] ***Lawrence J, Reed C.*** Mining Argumentative Structure from Natural Language text using Automatically Generated Premise-Conclusion Topic Models. In: Argument Mining (Denmark), proc. of the 4th Workshop. Association for Computational Linguistics, 2017: 39-48.
[9] ***Nguyen HV, Litman D.*** Context-aware argumentative relation mining. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Berlin). 2016; 1: Long Papers: 1127-1137.
[10] ***Niculae V, Park J, Cardie C.*** Argument mining with structured SVMs and RNNs. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics. 2017; 1: Long Papers: 985-995.
[11] ***Zasypkin AS, Pimenov IS, Salomatina NV***. The Combined Approach to Identifying Argumentation Structures in Short Scientific Papers. In: IEEE 24th International Conference of Young Professionals in Electron Devices and Materials (EDM). 2023: 1800-1805.
[12] ***Rahwan I, Reed C.*** The argument interchange format. In: G. Simari, I. Rahwan (ed.): Argumentation in Artificial Intelligence. Boston: Springer; 2009: 383-402.
[13] ***Walton D, Reed C, Macagno F.*** Argumentation schemes. Cambridge University Press; 2008. 443 p.
[14] ***Sidorova EA, Akhmadeeva IR, Zagorulko YuA, Sery AS, Shestakov VK.*** Research platform for the study of argumentation in popular science discourse [In Russian]. Ontology of designing. 2020; 10(4): 489-502.
[15] Scikit Learn Homepage. https://scikit-learn.org/stable/supervised_learning.html.
[16] Habr Homepage. https://habr.com/ru/articles.
[17] ***Pimenov IS.*** Analyzing Disagreements in Argumentation Annotation of Scientific Texts in Russian Language [In Russian]. In: NSU Vestnik. Series: Linguistics and Intercultural Communication. 2023; 21(2): 89-104.

About the authors

Natalia Vasilievna Salomatina (b.1958) graduated from the Novosibirsk State University in 1980, PhD (2009). She is a Senior Researcher of the Laboratory of Artificial Intelligence at the A.P. Ershov Institute of Informatics Systems (Novosibirsk, Russia). She is the author of more than 80 publications in the field of Automatic Speech Recognition, Sequence Analysis and Computational Linguistics. Author ID (RSCI): 5683; ORCID: 0000-0001-8412-9116; Author ID (Scopus): 57190173916; Researcher ID (WoS): G-3032-2019. *salomatina_nv@live.ru* ✉

Elena Anatolievna Sidorova (b. 1977) graduated from the Novosibirsk State University in 2000, PhD (2006). She is a Senior Researcher of the Laboratory of Artificial Intelligence at the A.P. Ershov Institute of Informatics Systems (Novosibirsk, Russia), and a Associate Professor at the Novosibirsk State University. She is a member of Russian and European Associations for Artificial Intelligence. Dr. Sidorova has more than 160 peer-reviewed publications in the field of Computational Linguistics, Intelligent System Development, Knowledge and Ontology Engineering. Author ID (RSCI): 146000; ORCID: 0000-0001-8731-3058; Author ID (Scopus): 41961707000; Researcher ID (WoS): K-2432-2018. *lsidorova@iis.nsk.su*.

Ivan Sergeevich Pimenov (b. 1997) graduated from the Novosibirsk State University master's program in 2020. He is a 2nd category programmer at the A.P. Ershov Institute of Informatics Systems SB RAS, and a postgraduate student at the Department of Fundamental and Applied Linguistics of the Novosibirsk State University. The list of his scientific works includes 15 works in the field of computational linguistics, including works on Argument Mining. ORCID: 0000-0001-5946-9469; Author ID (RSCI): 1164941. *pimenov.1330@yandex.ru*.

Received November 29, 2023. Revised January 18, 2024. Accepted February 2, 2024.