Электронный корпус научных текстов для задачи автоматического реферирования

Зизов Вадим Сергеевич
Кафедра математической кибернетики, Москва
E-mail: vadim1221@hotmail.com

В общем случае задача реферирования текста состоит в выделении важной информации во входном документе и построении краткого реферата. Особенностью текстов научного стиля является стремление к нормированию речи, строгий отбор языковых средств, использование специфических терминов и сложных синтаксических конструкций. Реферирование научных документов является частной задачей реферирования длинных документов. Для улучшения и облегчения работы моделей, работающих с длинным вводом, были предложены различные подходы обработки больших данных, такие как применение иерархического кодера совместно с декодером для обработки информации на уровне дискурса [1], отдельное реферирование частей или глав статей [2] и другие.

Отличительной чертой научных статей является наличие аннотации. Зачастую задача реферирования документа в научном стиле сводится к задаче написания аннотации, и наоборот, наличие корпуса статей с аннотациями помогает решить задачу реферирования в общем случае. Вместе с тем, недостаток корпусов научных текстов на русском языке пагубно сказывается на решении данной задачи. Отдельным направлением исследований является аспектный анализ текстов, сосредоточенный на выделении направлений мысли, таких как цель исследований, результат или метод.

Основы индикаторного подхода к извлечению информации из текста были заложены ещё в 70-е годы двадцатого века. Извлечение онтологической информации из научных текстов в [3] основано на индикаторном методе, в основе которого лежит обнаружение в тексте специфических подсказок в виде различных словесных клише: «в настоящей работе», «целью. . . является», «проведённое исследование» и т.п. Эти клише являются индикаторами того или иного аспекта содержания текста. В последнее время появились исследования, направленные на выявление шаблонных индикаторов структуры дискурса, например, в [4] описываются основные модели образования свободных конструкций, сигнализирующих о наличии причинно-следственного риторического отношения в дискурсе.

Успешные применения аспектного анализа для реферирования текстов в основном концентрировались на обобщении мнений из онлайн-обзоров [5], выделении аргументации и анализе новостных статей [6]. В [7] набор данных для автоматического реферирования научных текстов с аспектным анализом был собран с использованием структурированного описания научных статей, извлечённых из базы данных Emerald (www.emerald.com). Он покрывает широкий спектр предметных областей, в основном включая в себя маркетинг, менеджмент, образование и экономику.

Шаблонные конструкции, описывающие классы языковых выражений, применяются в различных задачах автоматической обработки текста. В зависимости от типа учитываемой в конструкции языковой информации, применяемые в различных работах шаблоны подразделяются на лексические, грамматические, лексико-грамматические и лексико-синтаксические. В [8] предусмотрена возможность описания сложного шаблона с использованием именованных подшаблонов, например, грамматически согласованных именных групп. Шаблоны имеют самое широкое применение в приложениях, связанных с извлечением информации, таких как извлечение именованных сущностей, терминов, их связей, реферирование и т.п. В работе [9] описываются методы на основе лексико-синтаксических шаблонов и правил, разработанные для автоматического извлечения и отбора терминов

в предметный указатель научного текста, а также для выявления их подчинительных связей.

Собранный корпус состоит из научных статей, опубликованных в журнале «Информационно-управляющие системы» с открытом доступом (OA). Журнал отличается своими требованиями к аннотациям, практически в явном виде формулирующим индикаторные фразы, относящиеся к статье. Это издание включает в себя статьи по двум отраслям науки и по трём группам специальностей ВАК (1.2, 2.2, 2.3). В корпус включено 729 статей, имеющих одновременно название, полный текст на русском языке и аннотацию. 223 из них имеют выраженные в явном виде аспекты статьи. Среди них 627 записей содержат также ключевые слова статьи.

Основным предназначением корпуса является помощь в построении и настройке систем реферирования текстов в научной области. Тексты статей предобработаны и разобраны на токены, помогая избежать ошибок, связанных с автоматическим преобразованием pdf-документов, аналогично преобразованию [10]. Отдельно выделены токены (ENG_), содержащие текст на английском языке (GRC_), формулы (FLÆ), ссылки на другие источники (L_). Использование юникода помогает снизить вероятность непредумышленного совпадения с некоторой содержательной фразой, и необходимостью хранить обработанный текст в незашифрованном виде.

Ценность явного выделения аспектов заключается в том, что они помогают выделить главную мысль текста, и в некоторых случаях облегчают структурирование аннотации. В качестве аспектов используются ключевые фразы «Результаты:», «Значимость:», «Цель:», «Методы:», «Проблемы:».

Собранный корпус опубликован по адресу https://mk.cs.msu.ru/images/8/81/Информационно-управляющие_системы.tsv.zip

Источники и литература

1) Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

2) Gidiotis, Alexios and Grigorios Tsoumakas. “Structured Summarization of Academic Publications.” PKDD/ECML Workshops (2019).

3) Саломатина Н.В., Гусев В.Д. Автоматизация формирования индикаторных словарей и возможности их использования // Труды межд. конференции Диалог-2006 «Компьютерная лингвистика и интеллектуальные технологии», Бекасово, 31мая − 4 июня 2006 Москва. "Наука". С. 121−125.

4) Toldova S., Pisarevskaya D., Vasilyeva M., Kobozeva M. The cues for rhetorical relations in Russian: "Cause-Effect" relation in Russian Rhetorical Structure Treebank // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”. 2018 Issue 17

5) Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

6) Lea Frermann and Alexandre Klementiev. 2019. Inducing Document Structure for Aspect-based Summarization. In Proceedings of the 57th Annual Meeting of the

Association for Computational Linguistics, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

7) Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing Structure into Summaries: a Faceted Summarization Dataset for Long Scientific Documents. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 1080–1089, Online. Association for Computational Linguistics.
8) Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексикосинтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог ‘2007 – М.: Издательский центр РГГУ, 2007 С. 70–75.
9) Большакова Е.И., Иванов К.М. Выделение терминов и их связей для предметного указателя научного текста // Сборник трудов XVI Национальной конференции по искусственному интеллекту с международным участием (КИИ-2018). Т1. М.: РКП, 2018 С. 253–261.
10) Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney and Daniel S. Weld. “S2ORC: The Semantic Scholar Open Research Corpus.” Annual Meeting of the Association for Computational Linguistics (2020).