Инженерный вестник Дона, №12 (2022)
ivdon.ru/ru/magazine/archive/n12y2022/8106

Извлечение ключевых понятий и связей между ними из тематических текстов на русском языке

М.Е. Денисов, А.М. Катышев, О.А. Сычев, А.В. Аникин

Волгоградский государственный технический университет, Волгоград

Аннотация: В работе рассматриваются подходы к решению таких задач обработки естественного языка, как извлечение ключевых понятий или терминов, а также семантических связей между ними для построения IT-решений на основе данных. Тема работы актуальна ввиду постоянного роста объёмов слабо структурированного и неструктурированного текста в электронном формате. Извлечённая информация может быть использована для улучшения многих процессов: автоматическое тегирование, оптимизация поиска по контенту, построение облаков слов и навигации; кроме того, для создания черновых версий словарей, тезаурусов и даже базы для экспертных систем.
Ключевые слова: обработка естественного языка, термин, лемма, семантическая связь, статистическая обработка, машинное обучение, word2vec.

Введение

Автоматическое извлечение из необработанных текстов ключевых понятий или концептов, а также семантических взаимосвязей между ними являются одними из актуальных задач обработки естественного языка, которые находят применение в задачах построения словарей, предметных тезаурусов и онтологий [1, 2]. В то время, как для английского и ряда европейских языков эти задачи решаются существующими средствами [3] с приемлемым качеством, для русского языка эта задача гораздо сложнее из-за его синтаксической гибкости – флективности, а также из-за большого лексического разнообразия. Кроме того, русский язык поддерживается лишь несколькими программными библиотеками широкого профиля, которые не предоставляют указанных возможностей. Таким образом, целесообразно разработать практические методы извлечения понятий и отношений с учётом особенностей русского языка.

Подход к извлечению терминов с использованием структуры текста

Большинство существующих метрик для ранжирования терминов [4] текста (TF-IDF, C-Value и др.) опираются на статистику по большой

коллекции документов [5, 6], сравнивая среднюю частоту встречаемости слова с частотой в исследуемом документе или тексте [7]. Без коллекции тематических текстов этих метрики непригодны, и возникает необходимость в методе, который использует для работы только данный текст.

При наличии достаточно длинного текста, как книга или учебник, предполагается естественное деление его на тематические разделы [8]. Совокупность разделов могла бы играть роль объединённой коллекции текстов для вычисления TF-IDF и т. п. Однако, здесь возникает проблема адекватного определения границ разделов. Пунктуация не является надёжным критерием из-за разнообразия стилей оформления; кроме того, в пределах одного раздела могут обсуждаться разные темы.

Позиции вхождений термина можно рассматривать по тексту в целом и работать со сглаженной плотностью встречаемости. Каждое слово или фраза имеет свой профиль встречаемости, который определяется количеством вхождений слова в каждую позицию текста. За элементарный элемент профиля встречаемости взято предложение как минимальная неделимая часть текста. Посчитав, сколько раз слово встречается в каждом предложении, предлагается перейти к сглаженному представлению профиля и оценить степень его неровности, вычислив среднеквадратическое отклонение.

Схема сжатия профиля встречаемости до любого требуемого размера:

а) для каждого предложения в тексте рассчитывается его вес, равный числу слов в нем;

б) для заданного слова вычисляется т.н. «сырой» профиль, длиной по количеству предложений;

в) затем методом суммирования отдельных элементов «сырой» профиль сжимается до указанного числа фрагментов, стремясь сохранить равномерное распределения веса в каждом объединённом фрагменте. Число

объединённых фрагментов не должно превышать 50–80, в то же время в каждом фрагменте должно быть значимое количество слов – не менее 50.

В русскоязычной литературе нередки термины, состоящие из нескольких слов, которые всегда употребляются вместе и означают единое понятие, например «среднеквадратическое отклонение» или «хи квадрат». Необходимо учитывать различные формы числа и падежа, в которых встречаются термины, при этом могут меняться окончания как отдельных слов, так и у всех слов словосочетания. Для решения этой проблемы принято приводить слова в некоторую «нормальную форму», в которой нет различий по числу и падежу. Для английского языка популярен стемминг – отсекание окончания от основы слова – «стема», а для русского больше подходит лемматизация – приведение слова к лемме – начальной форме (единственное число, именительный падеж).

Однако ни стемминг, ни лемматизация не защищены от проблемы неоднозначности, которая возникает при таком преобразовании: одна и та же буквенная запись может относиться к разным словам, порой совершенно далёким по смыслу, например, слово «полю» без контекста может означать иметь любую из следующих лемм: «поле», «Поля», «Поль», «полоть». Применяемая модель анализа текста рассматривает каждое слово в отдельности и не даёт информации о направлении разрешения подобных неоднозначностей. Пролить свет на настоящий смысл могут другие употребления этого слова в другой форме. Поэтому было принято решение разработать новую схему сравнения лемм, в которой используются все предполагаемые леммы слова. Согласно предлагаемой схеме, цепочки слов «совпадают по лемме», если длины цепочек в словах совпадают, и каждая пара соответствующих слов имеет хотя бы одну общую лемму.

В свете этого определения, например, цепочки «поле класса» и «полем классов» совпадают, так как обе состоят их двух слов, и каждая пара слов имеет пересекающиеся леммы: «поле» / «поль» и «класс».

Основные этапы процесса извлечения терминов включают:

а) токенизация – исправление опечаток и ошибок форматирования, разбивка массива символов на слова, фильтрация не-слов, удаление стоп-слов, анализ пунктуации для сегментации слов на фразы и предложения;

б) разбиение входного текста на фрагменты в соответствии с заданным числом фрагментов, получение профилей встречаемости слов в тексте;

в) поиск цепочек (существительных и именных групп) – кандидатов в термины – и слияние их различных форм;

г) расчёт значимости каждой цепочки-кандидата, равной произведению стандартного отклонения профиля встречаемости на число слов в цепочке;

д) определение терминов предметной области – сортировка терминов по убыванию рассчитанной значимости и взятие N лучших терминов.

Результаты экспериментов показывают применимость описанного метода на практике. Так, из книги, посвящённой паттернам проектирования [9], из первых 10 найденных терминов пять оказались названиями паттернов, описанных в этой книге, а остальные – релевантными понятиями по теме.

Подходы к извлечению семантических связей

Существует несколько подходов к задачам анализа подобного рода: на основе паттернов (или шаблонов), статистический (подсчёт встречаемости, использование нейросетей и т. п.), и различные эвристические методы.

Паттерны представляют собой одну из самых старых форм технического анализа. Чаще всего паттерн описывает устойчивую конструкцию с подстановочными местами для слов определённой части речи и/или стоящих определённой форме. Грамотно составленные паттерны могут обеспечить высокую точность извлечения, но для обеспечения полноты поиска нужно произвести большую работу по составлению паттернов, не внеся при этом противоречий между ними.

В противоположность, методы, основанные на статистике, требуют немного (или вообще не требуют) входных размеченных данных, и только выигрывают при увеличении обучающей репрезентативной выборки. Они мало зависят от контекста применения, например языка текста (русский, английский). Минусом этой группы методов является низкая начальная точность на выборках небольшого и среднего объёма, которую нередко требуется повышать параллельным применением других методов.

При непосредственном анализе статистических признаков важно угадать с выбором метрик, адекватно характеризующих рассматриваемые критерии предмета исследования. Нейросетевой подход позволяет переложить часть задачи выбора подходящих метрик на сам процесс обучения нейросети, однако подготовка данных специального формата, выбор архитектуры и размерностей нейросети в полном объёме остаётся заботой исследователей [10].

Благодаря особому свойству «приспосабливаться» способность нейросетей решать нечёткие задачи классификации и регрессии находит всё большее количество приложений в самых разных областях. Так, на нейросетевом принципе построен инструмент Word2Vec – разработанная в 2013 г. в Google модель дистрибутивного представления слов в векторном пространстве большой размерности. За девять лет существования этой технологии разработано множество эффективных методов извлечения

полезной информации разного рода (лингвистической, семантической и другой) из заранее подготовленных векторных моделей.

Заключение

В работе рассмотрены задачи автоматического извлечения терминов и семантических связей между понятиями, а также связанные с этим проблемы. Предложены подходы: к поиску терминов – на основе статистики без использования посторонних текстов, что облегчает переносимость подхода на другие языки; к извлечению отношений – на основе машинного обучения, с использованием языковых моделей, заранее подготовленных из общеязыковых корпусов текстов.

Литература

1. Добров Б. В. Онтологии и тезаурусы: модели, инструменты, приложения. Москва: Бином. Лаборатория знаний. 2009. 173 с.

2. Платонов А. В., Полещук Е. А. Методы автоматического построения онтологий // Программные продукты и системы. 2016. №2 (114). С. 47–52.

3. Anikin A., Litovkin D., Sarkisova E., Petrova T., Kultsova M. Ontology-based approach to deciison-making support of conceptual domain models creating and using in learning and scientific research. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012074.

4. Мякшин К. А. Разнообразие подходов к определению понятия «термин» // Альманах современной науки и образования. 2009. № 8-2. С. 109–111. URL: gramota.net/materials/1/2009/8-2/47.html

5. Frantzi K., Ananiadou S., Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method // International Journal on Digital Libraries. 2000. Vol. 3, № 2. P. 115–130.

6. Григорьева Е. Г., Клячин В. А. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник

Волгоградского государственного университета. Серия 2: Языкознание. 2017. Т. 16, № 2. С. 58–67. URL: doi.org/10.15688/jvolsu2.2017.2.6

7. Шереметьева С. О., Осминин П. Г. Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2015. Т. 12, № 1. С. 76–81. URL: elibrary.ru/item.asp?id=23057410.

8. Федоренко Д. Г., Астраханцев Н. А. Автоматическое извлечение новых концептов предметно-специфичных терминов // Труды Института системного программирования РАН. 2013. Т. 25. С. 167–178.

9. Швец А. Погружение в паттерны проектирования. 2018. 406 с. URL: refactoring.guru/ru/design-patterns/book (дата обращения 11.06.2019).

10. Anikin A., Katyshev A., Denisov M., Smirnov V., Litovkin D. Using Online Update of Distributional Semantics Models for Desicion-Making Support for Concepts Extraction in the Domain Ontology Learning Task. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012073.

References

1. Dobrov B. V. et.al. Ontologii i tezaurusy: modeli, instrumenty, prilozheniya [Ontologies and thesauri: models, tools, applications]. Moskva: Binom. Laboratoriya znaniy. 2009. 173 p.

2. Platonov A.V., Poleshchuk E. A. Metody avtomaticheskogo postroyeniya ontologiy. Programmnyye produkty i sistemy. 2016. №2 (114). P. 47–52.

3. Anikin A., Litovkin D., Sarkisova E., Petrova T., Kultsova M. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012074.

4. Myakshin K. A. Raznoobraziye podkhodov k opredeleniyu ponyatiya termin. Al′manakh sovremennoy nauki i obrazovaniya. 2009. № 8-2. P. 109–111. URL: gramota.net/materials/1/2009/8-2/47.html

5. Frantzi K., Ananiadou S., Mima H. International Journal on Digital Libraries. 2000. Vol. 3, № 2. P. 115–130.

6. Grigor′yeva E. G., Klyachin V. A. Vestnik Volgogradskogo gosudarstvennogo universiteta. 2017. Vol. 16, № 2. P. 58–67. URL: doi.org/10.15688/jvolsu2.2017.2.6

7. Sheremet′yeva S. O., Osminin P. G. Vestnik YUUrGU. 2015. Vol. 12, № 1. P. 76–81. URL: elibrary.ru/item.asp?id=23057410.

8. Fedorenko D. G., Astrakhantsev N. A. Avtomaticheskoye izvlecheniye novykh kontseptov predmetno-spetsifichnykh terminov. Trudy Instituta sistemnogo programmirovaniya RAN. 2013. Vol 25. P. 167–178.

9. Shvets A. Pogruzheniye v patterny proyektirovaniya [Dive into design patterns]. 2018. 406 c. URL: refactoring.guru/ru/design-patterns/book (access date: 11.06.2019).

10. Anikin A., Katyshev A., Denisov M., Smirnov V., Litovkin D. IOP Conf. Series: Materials Science and Engineering. 2019. №483. DOI: 10.1088/1757-899X/483/1/012073.