Инженерный вестник Дона, №2(2025)
ivdon.ru/ru/magazine/archive/n2y2025/9823

Сравнительный анализ методов извлечения знаний из текстов для построения онтологий

А.И. Егунова, Р.С. Комаров, С.А. Федосин,

С. Д. Шибайкин, Р. А. Жарков

Национальный исследовательский Мордовский государственный университет им. Н.П. Огарева, г. Саранск

Аннотация: Данная статья посвящена сравнительному анализу методов извлечения из текстов знаний, используемых для построения онтологий. Обозреваются разнообразные подходы извлечения, такие, как лексические, статистические, методы машинного обучения и глубокого обучения, а также онтологически ориентированные методы. В результате исследования сформулированы рекомендации по выбору наиболее эффективных методов в зависимости от специфики задачи и типа обрабатываемых данных. Особое внимание уделено рассмотрению современных методов, таких, как глубокое обучение и использование трансформеров, которые обеспечивают высокую точность анализа и позволяют учитывать сложные зависимости в текстах. Эти методы не только ускоряют процесс обработки данных, но и открывают новые возможности для автоматизации анализа информации. Описаны подходы к созданию онтологий с использованием специализированных языков и инструментов, таких, как OWL и Protégé, что делает результаты работы применимыми в широком спектре задач, включая управление данными, информационный поиск и разработку интеллектуальных систем.
Ключевые слова: онтология, извлечение знаний, классификация текстов, именованные сущности, машинное обучение, семантический анализ, модель.

Введение

Извлечение знаний из текстовой информации стало ключевым направлением исследований в таких областях, как машинный перевод, семантический веб и искусственный интеллект. Разработка онтологий способствует упорядочиванию знаний и улучшает взаимодействие между различными информационными системами. Эта работа сосредоточена на исследовании различных подходов к извлечению знаний и их применении в процессе создания онтологий.

При формировании онтологий выделяются несколько основных компонентов: классы, свойства, экземпляры, аксессоры и ограничения [1].

Классы (или типы) представляют собой основные категории или группы, представляющие общие концепции в конкретной предметной области, например, «Человек», «Автомобиль».

Свойства в онтологии выступают в качестве характеристик классов или описывают взаимосвязи между классами. Выделяется два типа свойств: датированные и объектные.

Датированные свойства (*data properties*) принимают значения в пределах заданного диапазона, например, возраст или высота.

Объектные свойства (*object properties*) определяют отношения между экземплярами классов, например, «владеет» или «работает в».

Экземплярами в онтологии являются конкретные объекты или элементы, которые относятся к классам, например, «Иван» как экземпляр класса «Человек».

Аксессоры и ограничения задают правила использования свойств, например, ограничивая значение свойства «возраст» только положительными числами [2].

Для извлечения ключевых компонентов онтологий можно эффективно использовать автоматизированные методы извлечения знаний из текстов. Эти подходы позволяют обрабатывать большие объемы данных и выявлять структуры и связи, которые могут быть сложными для обнаружения вручную. Современные технологии обработки естественного языка (*NLP*) и машинного обучения позволяют системам автоматически классифицировать и определять сущности и их характеристики, что значительно ускоряет процесс формирования и расширения онтологий [3]. Кроме того, автоматизация процесса извлечения знаний снижает риск ошибок, присущих человеческому анализу, и обеспечивает более единообразное и объективное извлечение информации.

Среди основных методов извлечения знаний из текстов выделяют лексические, статистические, методы машинного обучения, глубокое обучение и онтологически ориентированные подходы. Для удобства анализа эти методы были классифицированы по ряду критериев, включая их эффективность применения, существующие ограничения и тип данных, с которыми они способны работать.

Сравнение методов извлечения знаний

Лексические методы анализа текста основаны на предопределенных правилах и шаблонах, отражающих структуру языка.

Методы распознавание именованных сущностей (*NER*) включают определение и классификация именованных сущностей в тексте, таких как имена, организации и места. Для этого применяются правила (например, регулярные выражения) и предобученные модели, а также синтаксический анализ, включающий в себя разделение текстов на составляющие (слова и предложения) с помощью формальных грамматик [4].

Использование регулярных выражений и предобученных моделей позволяет автоматизировать процесс извлечения информации, что полезно для анализа большого объема данных, например, в новостных лентах или социальных медиа [5].

Синтаксический анализ дает возможность выявлять сложные структурные связи в тексте, что важно для понимания контекста и для задач, связанных с машинным переводом или поиском информации.

Примером использования может быть извлечение названий крупных компаний (например, «*Apple*», «*Samsung*») для анализа их упоминаний в медийном пространстве.

Статистические методы анализируют данные для выявления закономерностей и важных характеристик.

Одним из основных статистических методов является *TF-IDF (Term Frequency-Inverse Document Frequency)*.

TF (Частота термина) – количество вхождений термина в документе, вычисляемая по формуле:

$$TF(t) = \frac{n_t}{\sum_k n_t}, \quad (1)$$

где n_t – число вхождений слова t в конкретном документе; $\sum_k n_k$ – общее число слов в данном документе.

IDF (Обратная частота документа) – логарифмическая мера редкости термина, вычисляемая по формуле:

$$IDF(t) = \log\left(\frac{N}{n_t}\right), \quad (2)$$

где N – общее количество документов; n_t – количество документов, содержащих термин t.

Произведение *TF* и *IDF* позволяет выявить важность термина относительно документа и всего корпуса данных [6].

Методы машинного обучения обучают модели на исторических данных для автоматизации принимаемых решений. Примером данного метода может быть наивный байесовский классификатор. Он основан на применении теоремы Байеса с предположением о независимости признаков [7]. Каждому классу назначается вероятность, и выбирается класс с максимальной вероятностью. Классификатор использует теорему Байеса для вычисления вероятности принадлежности к классу:

$$P(C|X) = \frac{P(X|C) * P(C)}{P(X)}, \quad (3)$$

где P(C|X) – апостериорная вероятность (вероятность класса C, если известные признаки X); P(C|X) – правдоподобие (вероятность наблюдать признаки *X*, если класс *C* истинный); *P(C)* – априорная вероятность класса *C*;

$P(X)$ – нормализующий фактор (вероятность наблюдать признаки X, независимо от класса).

Наивное предположение означает, что все признаки X_i независимы, т.е.:

$$P(X|C) = P(X_1|C) * P(X_2|C) * ... * P(X_n|C), \quad (4)$$

Таким образом, формула для апостериорной вероятности будет выглядеть так:

$$P(C|X) = \frac{P(C) * P(X_1|C) * P(X_2|C) * ... * P(X_n|C)}{P(X)} \quad (5)$$

Для классификации объекта выбирается класс, соответствующий максимальной апостериорной вероятности:

$$C_{предсказанный} = ag \max_C P(C|X), \quad (6)$$

Глубокие нейронные сети, например, модели *BERT*, достигли высоких результатов в обработке естественного языка.

BERT (Bidirectional Encoder Representations from Transformers) использует механизм внимания для обучения контекстуального представления слов [8]. Модель обучается на задачах предсказания слов в предложениях с использованием двунаправленной контекстной информации [9].

Трансформеры – архитектура, которая позволяет моделям захватывать длинные зависимости в тексте путем внимательной агрегации информации из различных частей входных данных [10].

Пример использования трансформеров на языке программирования Python представлен на рис. 1.

```
from transformers import pipeline
nlp = pipeline("question-answering")
result = nlp(question="What causes the rise in ocean levels?",
             context="Climate change is causing ocean levels to rise.",
             inputs="")
print(result['answer'])
```

Рис. 1. – Программный код использования трансформеров

Модель возвращает «*Climate change*», что демонстрирует ее способность понимать контекст и извлекать ответы.

Еще одним способом извлечения знаний является онтологически ориентированные методы. Онтологии создаются на основе формальных языков, таких как *OWL (Web Ontology Language)*, и используют семантические технологии для описания объектов и их свойств. По запросам к таким онтологиям можно извлекать структурированную информацию с помощью языка запросов *SPARQL* [11].

После создания онтологии в *Protégé* можно выполнить запрос: *SELECT ?x WHERE { ?x rdf:type :Organism . },* который вернет всех организмов, представленных в онтологии, что позволяет проводить более сложный анализ данных.

Результаты исследования

Для более наглядного представления характеристик и применения различных методов анализа текста ниже приведена сравнительная таблица №1. Она включает основные преимущества и недостатки каждого подхода, а также их основные области использования. Представленные методы имеют свои уникальные особенности и области применения, что делает их подходящими для различных задач обработки текстов. Важно учитывать как сильные, так и слабые стороны каждого подхода при выборе инструмента для анализа данных. Данная таблица помогает систематизировать ключевые аспекты методов, облегчая их сравнение и выбор. Эти методы могут быть объединены и адаптированы для решения комплексных задач, требующих разнообразных подходов к анализу текстовой информации. Такой интегративный подход позволяет получить более полное представление о данных и повысить точность результатов.

Таблица №1.

Сравнительный анализ методов

Метод	Преимущества	Недостатки	Применение
Лексические методы	Простота реализации, высокая интерпретируемость	Ограниченная гибкость, необходимость словарей	Извлечение простых сущностей и фактов из структурированных текстов
Статистические методы	Способность обрабатывать большие объемы данных	Чувствительность к шуму в данных	Извлечение ключевых слов и фраз из неструктурированных текстов
Методы машинного обучения	Высокая точность, адаптивность	Требуют больших объемов размеченных данных	Классификация текстов и извлечение отношений между сущностями
Глубокое обучение	Способность обрабатывать сложные зависимости	Высокие вычислительные затраты	Извлечение знаний из больших текстовых корпусов, обработка естественного языка
Онтологически ориентированные подходы	Сильная структурированность и семантическая связность	Зависимость от качества существующих онтологий	Создание и расширение онтологий на основе имеющихся данных

По результатам анализа можно утверждать, что наиболее эффективными являются методы глубокого обучения и машинного обучения при обработке больших объемов текстовых данных. Однако лексические и статистические методы могут быть эффективны на начальных этапах анализа или для конкретных задач, где точность не является критичной.

Заключение

Среди рассмотренных методов для извлечения знаний и построения предметных онтологий из текста наиболее подходящими являются методы глубокого обучения, особенно модели типа BERT. Эти методы демонстрируют высокую эффективность в обработке естественного языка благодаря способности учитывать контекст и выявлять сложные зависимости. Хотя глубокое обучение требует значительных ресурсов и размеченных данных, его преимущества в точности и способности обрабатывать большие объемы данных делают его особенно ценным инструментом для создания и расширения онтологий. Таким образом, использование BERT и аналогичных моделей представляется наиболее перспективным направлением для извлечения знаний из текстовых источников.

Литература

1. Загорулько, Ю. А., Боровикова О.И., Загорулько Г.Б. Применение паттернов онтологического проектирования при разработке онтологий научных предметных областей // Аналитика и управление данными в областях с интенсивным использованием данных: Сборник научных трудов XIX Международной конференции DAMDID / RCDL'2017, Москва, 10–13 октября 2017 года / Под ред. Л.А. Калиниченко, Я. Манолопулос, Н.А. Скворцова, В.А. Сухомлина. – Москва: Федеральный исследовательский центр "Информатика и управление" Российской академии наук, 2017. – С. 332-339. – EDN XQVAWL.

2. Lamy Jean-Baptiste Ontologies with Python Programming OWL 2.0 Ontologies with Python and Owlready2. 1 st ed. 2021. pp. 89-93.

3. Герасимова С.В., Мандрик А.Д. Роль онтологий и семантических моделей в обеспечении связности и релевантности ответов диалоговых систем // Тенденции развития интернет и цифровой экономики: Труды VII

Международной научно-практической конференции, Симферополь-Сатера (Алушта), 30 мая – 01 2024 года. – Симферополь: ИП Зуева, 2024. – С. 124-125.

4. Бабина О.И., Панюков А.В. Байесовский классификатор именованных сущностей в тексте на естественном языке // Информационные технологии и системы: Труды Четвертой Международной научной конференции, Банное, 25 февраля 2015 года / Ответственные редакторы: Ю.С. Попков, А.В. Мельников. – Банное: Челябинский государственный университет, 2015. – С. 7-9.

5. Егунова А. И., Комаров Р.С., Вечканова Ю.С. [и др.] Анализ алгоритмов и решений для автоматической генерации подводок новостных статей в соцсетях с использованием искусственного интеллекта // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2023. – № 1. – С. 25-35. – DOI 10.24143/2072-9502-2023-1-25-35. – EDN EGRKGN.

6. Маслова, М. А., Дмитриев А.С., Холкин Д.О. Методы распознавания именованных сущностей в русском языке // Инженерный вестник Дона. – 2021. – № 7. URL: ivdon.ru/ru/magazine/archive/n7y2021/7066.

7. Челышев Э. А., Раскатова М. В., Мишин А. А., Щеголев П. Автоматическое реферирование текстов: обзор алгоритмов и подходов к оценке качества // Инженерный вестник Дона. – 2023. – № 12. URL: ivdon.ru/ru/magazine/archive/n12y2023/8871.

8. Вечканова Ю.С., Федосин С.А. Использование BERT-моделей естественных языков для управления нормативно-справочной информацией // L Огарёвские чтения: Материалы всероссийской с международным участием научной конференции. В 3-х частях, Саранск, 06 ноября – 11 2021 года / Отв. за выпуск А.М. Давыдкин, сост. Г.В. Терехина. Том Часть 1. –

Саранск: Национальный исследовательский Мордовский государственный университет им. Н.П. Огарёва, 2022. – С. 416-419.

9. Сайгин А.А., Плотникова Н.П. Векторизация нормативно - справочной информации с помощью модели нейронной сети BERT // Информационные технологии и математическое моделирование в управлении сложными системами. – 2021. – № 2(10). – С. 52-59.

10. Hirst G., Lin J., Nogueira R., Yates A. Pretrained Transformers for Text Ranking: BERT and beyond // Synthesis Lectures on Human Language Technologies. – 2021. – Vol. 14, No. 4. – pp. 1-325. – DOI 10.2200/S01123ED1V01Y202108HLT053. – EDN GQYVZK.

11. Дубинин В. Н., Дубинин А. В., Янг Ч.В., Вяткин Ч. В. Использование языка sparql в онтологическом моделировании мультиагентных систем в семантическом WEB // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2020. – № 1(53). – С. 4-18.

References

1. Zagorulko Yu., Borovikova O., Zagorulko G. / Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017) Moscow. Russia. October 9-13, 2017. Leonid Kalinichenko, Yannis Manolopoulos, Nikolay Skvortsov, Vladimir Sukhomlin (Eds.). pp. 258-265. (in Russian).

2. Lamy Jean-Baptiste Ontologies with Python Programming OWL 2.0 Ontologies with Python and Owlready2. 1 st ed. 2021. pp. 89-39.

3. Gerasimova S.V., Mandrik A.D. Tendentsii razvitiya internet i tsifrovoy ekonomiki: Trudy VII Mezhdunarodnoy nauchno-prakticheskoy konferentsii, Simferopol'-Satera (Alushta), 30 maya – 01 2024 goda. – Simferopol': IP Zuyeva, 2024, pp. 124-125.

4. Babina O.I., Panyukov A.V. Bayyesovskiy klassifikator imenovannykh sushchnostey v tekste na yestestvennom yazyke, Informatsionnyye tekhnologii i sistemy: Trudy Chetvertoy Mezhdunarodnoy nauchnoy konferentsii, Bannoye, 25 fevralya 2015 goda. Otvetstvennyye redaktory: YU.S. Popkov, A.V. Mel'nikov. Bannoye: Chelyabinskiy gosudarstvennyy universitet, 2015, pp. 7-9.

5. Yegunova A.I., Komarov R.S., Vechkanova YU.S., Sidorov D.P., Shibaykin S.D., Nikulin V.V. Vestnik Astrakhanskogo gosudarstvennogo tekhnicheskogo universiteta. Seriya: Upravleniye, vychislitel'naya tekhnika i informatika. 2023, № 1, pp.25-35

6. Maslova M.A., Dmitriev A.S., Kholkin D.O. Inzhenernyi vestnik Dona, 2021, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2021/7066.

7. Chelyshev E.A., Raskatova M.V., Mishin A.A., Shchegolev P.V. Inzhenernyi vestnik Dona, 2023. № 12. URL: ivdon.ru/ru/magazine/archive/n12y2023/8871.

8. Vechkanova YU.S., Fedosin S.A. L Ogarovskiye chteniya: Materialy vserossiyskoy s mezhdunarodnym uchastiyem nauchnoy konferentsii. V 3-kh chastyakh, Saransk, 06 noyabrya – 11 2021 goda. Otv. za vypusk A.M. Davydkin, sost. G.V. Terekhina. Tom Chast' 1. Saransk: Natsional'nyy issledovatel'skiy Mordovskiy gosudarstvennyy universitet im. N.P. Ogarova, 2022, pp. 416-419.

9. Saygin A.A., Plotnikova N.P. Informatsionnyye tekhnologii i matematicheskoye modelirovaniye v upravlenii slozhnymi sistemami. 2021, № 2(10), pp.52-59.

10. Hirst G., Lin J., Nogueira R., Yates A. Synthesis Lectures on Human Language Technologies. 2021. Vol. 14, №4. pp. 1-325. p DOI 10.2200/S01123ED1V01Y202108HLT053. EDN GQYVZK.

11. Dubinin V. N., Dubinin A. V., Yang CH.V., Vyatkin CH. V. Izvestiya vysshikh uchebnykh zavedeniy. Povolzhskiy region. Tekhnicheskiye nauki. 2020, № 1(53), pp. 4-18.

Дата поступления: 17.12.2024

Дата публикации: 26.01.2025