Якушева К.И. Способ извлечения утверждений из неформализованного текста на основе онтологий // Международный журнал информационных технологий и энергоэффективности. – 2019. – Т. 4 № 3(13) с. –3-9

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.912

СПОСОБ ИЗВЛЕЧЕНИЯ УТВЕРЖДЕНИЙ ИЗ НЕФОРМАЛИЗОВАННОГО ТЕКСТА НА ОСНОВЕ ОНТОЛОГИЙ

Якушева К.И.

Филиал федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет МЭИ» в г. Смоленске, Россия (214013, г. Смоленск, Энергетический проезд, дом 1); e-mail: www.roksana@mail.ru

В данной статье предлагается способ извлечения утверждений на основе онтологии для повышения качества извлечения. Суть способа состоит в применении онтологии предметной области на этапе извлечения отношений. Способ извлечения утверждений базируется на подходе по правилам. Для решения задачи извлечения помимо онтологий применяются лингвистические правила и шаблоны. Способ включает в себя такие этапы анализа текста как: графематический, морфологический, синтаксический. Также в статье представлена методика, позволяющая оценить качество извлечения утверждений из неформализованного текста по двум основным показателям: точности и полноте извлечения. Для объединения этих показателей предложено использовать F-меру. Методика оценки эффективности основана на официальных метриках РОМИП (Российский семинар по оценке методов информационного поиска).

Ключевые слова: извлечение утверждений, методика оценки эффективности, оценка точности и полноты, извлечение утверждений на основе онтологий.

A METHOD OF EXTRACTING STATEMENTS FROM UNSTRUCTURED TEXT USING ONTOLOGY

Yakusheva K.I.

Smolensk Branch of Federal state budgetary educational institution of higher education "National research University Moscow power engineering Institute", Russia (214013, Smolensk, Energeticheski proezd, 1); e-mail: www.roksana@mail.ru

This article proposes a method of extracting assertions based on ontology to improve the quality of extraction. The essence of the method consists in applying the ontology of the subject area at the stage of extracting relationships. The method of extracting assertions is based on a rule-based approach. In addition to ontologies, linguistic rules and patterns are used to solve the extraction problem. The method includes the following stages of text analysis. The article also presents a method for assessing the quality of extracting statements from a non-formalized text using two main indicators: accuracy and completeness of extraction. To combine these indicators, it is proposed to use the F-measure. The methodology for evaluating the effectiveness is based on the official ROMIP metrics (Russian seminar on the evaluation of information retrieval methods).

Keywords: extracting assertions, methods for evaluating effectiveness, evaluating accuracy and completeness, extracting assertions based on ontologies.

Задачи извлечения информации из текстов на сегодняшний день являются достаточно важными в компьютерной лингвистике. Можно выделить как подзадачу извлечение утверждений. Эта подзадача актуальна в тех случаях, когда есть потребность извлекать информацию, к примеру, для пополнения онтологий. Исходя из существующих подходов к извлечению информации и принимая во внимание необходимость извлечения именно утверждений в определенной тематике, можно сделать вывод о том, что наиболее подходящим является подход, основанный на правилах. Также учитывая все особенности этого подхода, можно утверждать, что для достижения более высокой точности и полноты извлечения оптимально использовать сочетание таких средств как лингвистические шаблоны и онтология предметной области. Это позволит компенсировать невозможность предусмотреть все варианты шаблонов и те случаи, когда шаблоны оказываются неэффективны.

Поэтому ставится цель:

- разработать способ и реализующий его алгоритм извлечения утверждений из неформализованного текста на основе онтологий, который будет более эффективен, чем классический подход, основанный на правилах;
- разработать методику оценки эффективности способа извлечения утверждений.

Способ извлечения утверждений из неформализованного текста на основе онтологий

К разрабатываемому способу предъявляются следующие требования:

- предлагаемый способ должен учитывать все этапы обработки текста: графематический, морфологический, синтаксический;
- разрабатываемы способ должен учитывать тематику предметной области (например, книжный каталог/магазин);
- способ извлечения утверждений должен предназначаться для текстов русского языка;
- разрабатываемы способ должен предполагать использование онтологии предметной области (книжный каталог/магазин);
- лингвистические шаблоны в способе извлечения утверждений должны основываться на заданной тематике и учитывать максимально возможное количество вариантов.

За основу способа берется метод извлечения фактов предложенный в [1]. Изменения относятся к 4 этапу, в который добавлен новый алгоритм, также изменен 5 и 6 этап, так как способы различаются извлекаемой информацией. Основным отличием можно назвать использование онтологии предметной области для извлечения отношений.

Предлагаемый способ извлечения утверждений из неформализованного текста состоит из следующих основных этапов.

Этап 1. Разбиение текста на абзацы и предложения.

Этап 2. Разбор каждого предложения на простые элементы: слова, пробелы, цифры, кавычки, знаки препинания и т.д.

Этап 3. Выявления сущностей, содержащихся в предложении: предикатов, объектов, предложенных групп, численных значений, дат и т.п.

Этап 4. Выявления связей между сущностями (по фактам или шаблонам).

Этап 5. Отбор отношений для формирования утверждений.

Этап 6. Проверка наличия неопределенности в указанных утверждениях.

Этап 7. Вывод полученных утверждений.

Далее более подробно раскрываются основные этапы.

Этап 1 Графематический анализ текста. На данном этапе с текстом производятся следующие действия.

– разбиение текста на графемы;

– выделение в исходном тексте абзацев, заголовков, примечаний;

– определение границ предложений.

На данном этапе могут возникнуть две связанные задачи:

1) определение, терминального знака препинания (под терминальными знаками понимается точка, восклицательный и вопросительный знаки) как границы предложения в данном контексте,

2) определение всех границ предложений в документе.

На данном этапе часто используются словари сокращений, которые помогают частично решить задачу определения конца предложения.

Выходные данные этого этапа являются входными для следующего и представляют размеченный текст. Возможно также табличное преставление, где каждый элемент таблицы – это отдельное предложение.

Этап 2 Предполагает разделение текста на еще более мелкие части: слова, пробелы, знаки препинания, цифры, скобки и т.д. Этот этап также, как и предыдущий можно отнести к графематическому анализу.

Выходные данные для этого этапа - таблица с отдельными словами, знаками препинания, цифрами и т.д.

Этот этап заканчивает подготовку текста к извлечению сущностей.

Этап 3 Осуществляет выявление сущностей. Первоначально необходимо определять имена собственные, числовые значения, названия в кавычках и т.п. После этого удаляем лишние пробелы. После чего происходит определение предикатов объектов (чаще всего глагол (причастие, деепричастие), а также предложных групп и дата/время.

На этом этапе используются дополнительные словари, например, словарь имен, фамилий, чисел и т.п.

Ключевым элементом этапа 3 разработанного метода извлечения утверждений являются правила поиска сущностей.

Предложено описывать каждое правило в виде функции определенного вида.

Функции необходимые к реализации разработанного метода:

1. функция для определения соответствия ли входного элемента правилу отбора сущностей;

2. функция поиска во множестве входных данных хотя бы одного элемента, удовлетворяющего правилу отбора сущностей;

3. функция перебора во множестве элементов входных данных, при этом одновременно все элементы должны соответствовать правилу отбора сущностей.

Для каждой из этих функций должны быть заданы три параметра: последовательность операций над данными для проверки соответствия, минимальное число совпадений и максимальное число совпадений. Для выполнения операций над данными каждая функция может вызывать другие функции.

Этап 4 Извлечение отношений между сущностями.

На данном этапе помимо правил, как было в изначальном методе, подключается онтологический подход. Система ищет совпадение слов из одного предложения и сравнивает с утверждениями из онтологии.

Важно, чтобы онтология была максимально наполнена и была релевантной наборам текстов, из которых будут извлекаться утверждения.

Шаблоны и правила, также помогают установить связь между сущностями, извлеченными в предыдущем этапе.

Основные сущности языка в обрабатываемых в текстах [2]:

- предикат (predicate) – множество элементов, которое состоит из глагола (причастие, деепричастие) или краткого прилагательного и дополняющих слов (наречий);
- дата/время (date/time) – множество элементов, обозначающих дату и (или) время;
- объект (object) – множество элементов, обозначающее одну сущность (явление, процесс) реального мира;
- предложная группа (prepositionalgroup) – множество элементов, которое состоит из предлога и объекта (объектов).

Основные отношения [3]:

- Объект – Предикат – Объект (ObjectPredicateObject) – множество элементов, обозначающее взаимодействие между несколькими объектами;
- Объект – Предикат – Свойство (ObjectPredicateProperty) – множество элементов, обозначающее свойство объекта.

Этап 5 Отбор отношений должен учитывать следующие критерии:

1) наличие важного ключевого предикативного слова (глагол, причастие, деепричастие, краткое прилагательное);
2) наличие важного ключевого слова-существительного;
3) наличие трех элементов в составе утверждения.

При соблюдении всех трех условий утверждение можно считать сформированным.

Этап 6 разработанного способа при наличии одинаковых утверждений, полученных из различных источников информации – проверка наличия неопределенности в указанных утверждениях. Возможен вариант полного совпадения утверждений, в этом случае один из вариантов отбрасывается автоматически.

Этап 7 Полученное множество утверждений группируется и представляется пользователю в виде отчетного документа. Возможно формирование текстового документа или вывод его на экран в рамках программного интерфейса.

Извлечение отношений происходит с помощью обращения к онтологии по средствам запросов на языке SparQL [4].

Методика оценки эффективности

Оценка эффективности системы извлечения утверждений базируется на официальных метриках РОМИП (Российский семинар по оценке методов информационного поиска) [5].

Для оценки качества работы системы применяются разные оценки, которые основываются на анализе результатов работы системы. При этом "самыс лучшим" алгоритмом можно считать тот, для которого выводы, сделанные системой, совпадают с мнением экспертов.

Большинство метрик, применяемых в современной оценке текстового поиска, опирается на отношении релевантности (принадлежности) найденного утверждения эталонному.

Метрики для неупорядоченного множества утверждений основаны на бинарной классификации документов «релевантен/не релевантен» по отношению к выбранному утверждению. Эти метрики основываются на матрице классификации (таблица 1), которая применятся в задачах поиска, извлечения фактов, событий и т.п. Применение метрики к утверждениям приемлемо, так как они сходны с фактами и включают в себя необходимые критерии оценки.

Таблица 1. Основные категории документов ответа системы

	Релевантны	Не релевантны
Найдено системой	a	b
Не найдено системой	c	d

Здесь, a — количество утверждений, найденных системой и релевантных с точки зрения экспертов; b — количество утверждений, найденных системой, но не релевантных с точки зрения экспертов; c — количество релевантных утверждений, не найденных системой; d — количество нерелевантных утверждений, не найденных системой.

Основными критериями, интересующим нас, являются точность и полнота.

Полнота (recall)

Полнота (recall) вычисляется как отношение найденных релевантных утверждений к общему количеству релевантных утверждений:

$$r = \frac{a}{a+c} \quad (1)$$

Полноту можно охарактеризовать как способность системы находить нужные пользователю утверждения, но она не учитывает количество нерелевантных утверждений, выдаваемых пользователю. Например, если полнота равна 50%, то это значит, что половина релевантных утверждений системой не найдена.

Точность (precision)

Точность (precision) вычисляется как отношение найденных релевантных утверждений к общему количеству найденных утверждений:

$$p = \frac{a}{a+b} \quad (2)$$

Точность можно охарактеризовать как способность системы выдавать в списке результатов только релевантные утверждения. Например, если точность равна 50%, то это значит, что среди найденных утверждений половина релевантных и половина – нерелевантных.

Иногда имеет смысл совместить эти две характеристики в одну, в таком случае можно использовать F-меру.

F-мера (F-measure)

F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного случая вычисляется по формуле:

$$F = \frac{2}{\frac{1}{p}+\frac{1}{r}} \quad (3)$$

Отметим основные свойства метрики F:

- $0 \le F \le 1$
- Если $p = 0$ или $r = 0$, то $F = 0$
- Если $p = r$, то $F = p = r$

- $\min(p, r) \leq F \leq \frac{p+r}{2}$

Предлагается использовать данную оценку эффективности таким образом.

1) Система должна обработать набор текстов, состоящих не менее чем из десяти текстов, для того чтобы избежать случайных ошибок.

2) Тексты предоставляются группе экспертов-оценщиков, они определяют эталонные утверждения, согласно предметной области.

3) Затем результаты экспертов-оценщиков сравниваются с результатами системы и делятся на 3 группы, согласно матрице классификации:

- найдены системой совпадают с эталонными утверждениями;
- не найдены системой эталонные утверждения;
- найдены системой ошибочно.

4) Подсчитав количественные показатели из предыдущего пункта находятся показатели полноты и точности.

5) Вычисляется F-мера, которая будет использоваться для оценки различных подходов.

6) Вычисляется относительное приращение показателей качества.

Для тестирования рекомендуется подбирать тексты таким образом, чтобы они не содержали опечаток и грамматических и пунктуационных ошибок, или предварительно проверять их, это повысит точность извлечения.

Оценка эффективности

Разработанные программные средства, основанные на предлагаемом способе, тестировались на наборе данных состоящих из 10 текстов по тематике предметной области. В качестве входных данных берутся рецензии пользователей книжного сервиса LiveLib [6]. Они представлены в виде текстового файла. Текст рецензий не структурирован, все авторские ошибки и опечатки сохраняются.

В таблице 1 приведены результаты оценки способа извлечения утверждений на основе онтологий со способом без использования онтологического подхода.

Таблица 1 – сравнение оценок эффективности.

	С использованием онтологического подхода	Без использования онтологического подхода
Точность (p)	0,83	0,75
Полнота (r)	0,65	0,44
F-мера	0,73	0,56

Таким образом показатели точности извлечения при использовании онтологии в среднем больше на 8%, а показатель полноты на 21%, если сравнивать общую эффективность извлечения утверждений из неформализованного текста на основе онтологий, то данный способ дают прирост эффективности в 17% по сравнению со способом, построенным на базе правил и лингвистических шаблонов.

В рамках статьи предложен способ извлечения утверждений на основе онтологий. Данный способ повышает качество извлекаемых утверждений из неформализованного текста в сравнении с подходом на правилах без использования онтологического подхода.

Также в статье приводиться методика оценки эффективности извлечения утверждений и предлагается для оценки описанного способа и представлены результаты оценки предложенного способа.

Список литературы

1. Grishman R., Information Extraction. In: The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds), Wiley-Blackwell, 2010, pp. 515-530.
2. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. – М.: НИУ ВШЭ, 2017, 296 с.
3. Sekine’s Extended Named Entity Hierarchy. URL: http://nlp.cs.nyu.edu/ene/.(дата обращения 21.05.2019).
4. Feldman R., Sanger J. (ed.). The text mining handbook: advanced approaches in analyzing unstructured data. — Cambridge University Press, 2007.
5. Gruber, T. R. A translation approach to portable ontology specification. // KnowledgeAcquisition.1993. Vol. 5. № 1. Pp. 199–220.
6. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии. – Москва, Издательство МГТУ имени Н.Э. Баумана, 2005. – 304 с.

References

1. Grishman R., Information Extraction. In: The Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox, and S. Lappin (Eds), Wiley-Blackwell, 2010, pp. 515-530.
2. Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. Automatic processing of natural language texts and data analysis. - M .: HSE, 2017, 296 p.
3. Sekine’s Extended Named Entity Hierarchy. URL: http://nlp.cs.nyu.edu/ene/.(referral date of May 21, 2019).
4. Feldman R., Sanger J. (ed.). The text mining handbook: advanced analysis in analyzing unstructured data. - Cambridge University Press, 2007.
5. Gruber, T. R. A translation approach to portable ontology specification. // KnowledgeAcquisition.1993. Vol. 5. No. 1. Pp. 199-220.
6. Bashmakov A.I., Bashmakov I.A. Intellectual information technologies. - Moscow, Publishing House of Moscow State Technical University named after NE Bauman, 2005. - 304 p.