УДК 004.912

**АНАЛИЗ ПОДХОДОВ И ЛИНГВИСТИЧЕСКИХ СРЕДСТВ РЕШЕНИЯ ЗАДАЧИ ИЗВЛЕЧЕНИЯ УТВЕРЖДЕНИЙ ИЗ НЕФОРМАЛИЗОВАННЫХ ТЕКСТОВ**

**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 статье рассматриваются основные виды извлекаемой информации. Описываются особенности извлечения из неформализованного текста именованных сущностей, отношений, фактов и событий. Приводиться анализ двух подходов: машинного обучения и подхода на правилах (инженерного подхода). Описаны плюсы и минусы данных подходов. Приводится обзор и анализ лингвистических средств, использующихся при подходе на правилах: лингвистических шаблонов, лингвистических правил и онтологий. Описаны достоинства и недостатки каждого из этих лингвистических средств. В статье представлен выводы о целесообразности использования онтологии для повышения точности и полноты извлечения утверждений из неформализованного текста.**

Ключевые слова: извлечение информации из текста, компьютерная лингвистика, подход на правилах, машинное обучение, онтологии предметной области, лингвистические шаблон, лингвистические привила.

**ANALYSIS OF APPROACHES AND LINGUISTIC MEANS OF SOLVING THE PROBLEM OF EXTRACTING ASSERTIONS FROM UNSTRUCTURED TEXTS**

**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*

**The article discusses the main types of extracted information. The features of extracting named entities, relations, facts and events from non-formalized text are described. An analysis of two approaches is given: machine learning and the rule-based approach (engineering approach). Pros and cons of these approaches are described. A review and analysis of the linguistic means used in the approach on the rules: linguistic patterns, linguistic rules and ontologies. The advantages and disadvantages of each of these linguistic means are described. The article presents conclusions about the appropriateness of using ontology to improve the accuracy and completeness of extracting statements from a non-formalized text.**

Keywords: Information Extraction, Text Mining, rule-based approach, machine learning, ontology, lexical approach.

Одной из значимых практических задач компьютерной лингвистики(КЛ), можно назвать задачу извлечение информации из текста *(Information Extraction)* [1],ее также можно отнести к направлению Text Mining. Это востребовано в решении задач экономической и производственной аналитики.

Решение этой задачи связано с выявлением в тесте на естественном языке (ЕЯ) определенных объектов – именованных сущностей (имен персоналий, географических названий, названий фирм и пр.), их отношений и связанных с ними фактов или событий. Также сюда можно отнести выявление утверждений. Утверждения очень похоже на факты и события, но строятся по строго определенной структуре и могут быть представлены тройкой, вида Субъект(Объект) – Предикат – Объект(Свойство). Пример выглядит таким образом: субъект — «трава», предикат — «имеет цвет», объект — «зеленый».

Основными этапами извлечение информации из текста можно назвать извлечение сущностей, извлечение отношений и извлечение фактов.

Целью данной статьи является анализ некоторых особенности извлечение сущностей и отношений, а также анализ основных подходов и некоторых лингвистических средств к решению данной задачи.

**Особенности извлечения сущностей из неформализованного текста**

Зачастую, текст из которого требуется извлекать данные слабо структурирован или не структурирован вовсе, поэтому в дальнейшем будем рассматривать случай извлечения, когда текст неструктурированный, его еще называют неформализованный.

Именованными сущностями принято называть объекты определенного типа [2]. Изначально выделяли:

* имена персоналий,
* название организаций,
* географические объекты.

На сегодняшний день к ним также можно добавить:

* даты и временные отрезки,
* номера телефонов,
* адреса,
* марки товаров,
* обозначения денежных единиц,
* ссылки на литературу,
* гены, белки, химические вещества.

Списка категорий (видов) именованных сущностей, который бы являлся общепринятым – нет, но существуют работы, в которых описываются попытки их создать – см. расширенную иерархию, состоящую из 200 видов [3].

Извлечение именованных сущностей связано с рядом сложностей.

1. Существует большое число разных сущностей и объектов, и их число постоянно увеличивается в связи с появлением новых.
2. Существует множество различных способов именования одной и той же сущности. Примером может служить William H. Gates, которого также можно именовать как Билл Гейтс, или владелец компании Microsoft.
3. Часто возникает проблема установления кореференции имен. Кореференция – это вид текстовой или синтаксической связности, при которой две или более номинативных (именных) групп называют один и тот же объект (референт). Примером кореференции можно считать ГАИ и ГИБДД – по сути обозначают один и тот же референт.
4. Большая зависимость от контекста. Именованные сущности могут относиться к разным видам (категориям)*.* Например, если мы говорим о надвигающиеся на центральную часть России грозы – речь о России идет как о географическом объекте. А если, говорим о вступлении России и ЕС – Россия выступает уже как страна.

Полное решение задачи извлечения именованной сущности в общем случае включает:

1. нахождение наименований сущности в тексте;
2. определение категории сущности;
3. связывание сущности с референтом (называемым лицом/объектом), если именованная сущность является именем собственным.

**Особенности извлечения отношений из неформализованного текста**

Следующей по сложности задачей можно назвать задачу извлечение отношений между объектами; обычно рассматриваются отношения только между двумя объектами. Это актуально для задачи извлечения утверждений, т.к. утверждение представляет собой связь двух сущностей и отношение между ними. Типы извлекаемых отношений зависят от прикладной задачи и предметной области текстов. Например, из новостных текстов можно получить информацию о том, кто какую должность занимает (отношение «занимать должность»), из научных статей по химии извлечь информацию о взаимодействии веществ (отношение «вступать в реакцию») и др.

Для выявления в тексте отношений так или иначе требуется привлечение информации о типичных конструкциях (контекстах) их выражения.

* атрибуты конкретных объектов. Например, атрибутами объекта – квартира (продажа/покупка), могут выступать: этаж, адрес, метраж, количество комнат, количество этажей в доме и т.д.;
* отношения (связи) конкретных объектов.

Можно выделить два вида отношений:

* общие: часть-целое, причина-следствие;
* зависящие от предметной области (ПрО) текста: работать\_в, быть\_учасником, вступать\_в\_организацию.

При извлечении учитываются типичные конструкции описания атрибутов и отношений. Сложность заключается в том, что отношения могут быть непостоянны.

**Особенности извлечения фактов и событий из неформализованного текста**

По своей сути извлечения фактов и событий объединяет в себе извлечение сущностей и отношений, поэтому можно сказать, что извлечение фактов самая сложная задача при извлечении информации в целом. Эта задача может использоваться для формирования новостных лент или служить основой для таких задач КЛ как, аннотирование и реферирование текста.

Сложность состоит в том, что в событие может участвовать несколько именованных сущностей, которые связаны определенным набором отношений. Например, событие «выдача кредита» фиксирует кто, кому, когда и в каком размере выдал кредит:

*Юго-Восточный банк Сбербанка выдал в 2014 году оборонным предприятиям кредиты на 17,8 млрд рублей.*

Можно заметить, что событие описывается определенным набором атрибутов (параметров) и их значений, в данном примере: заемщик, кредитор, сумма займа, время займа (может быть также добавлен атрибут срок возврата кредита). Такой набор образует так называемый семантический фрейм события [4]; значениями атрибутов выступают именованные сущности.

Особенности, которые стоит учитывать при извлечении фактов и событий:

1. событие или факт в тексте может быть выражено по-разному. Например, *Правительство РФ ответило американскому премьеру …* или *В Правительстве РФ ответили на обвинения американского премьера …*
2. часто сложно найти слово или словосочетание, которое выражает суть события.
3. могут встречаться слова, изменяющие суть (с частицами *почти, не*).
4. нередко необходимо слияние частичных описаний, полученных из разных предложений.

**Анализ существующих подходов к извлечению утверждений**

*Машинное обучение* – подход при котором требует большого объема вводных данных. Нужно создать максимальное покрытие лингвистической информацией обучающей выборки текстов: разметить всю морфологию, синтаксис, семантику.

Плюсы такого подхода заключаются в том, что он не требует ручного труда кроме создания размеченного корпуса. Не требуется составления лингвистических правил, шаблонов или онтологий. Если необходимо то, такая система легко перенастраивается и переобучается. Правила получаются более абстрактными.

К минусам можно отнести следующие. Инструменты для автоматической разметки русскоязычных текстов очень развиты, а существующие не всегда легкодоступны. Корпуса текстов должны быть достаточно объемными, правильно, однообразно и полностью размечены, что является достаточно трудоемким процессом. Кроме этого, в них трудно отследить ошибки, и точечно ее исправить.

*Подход, основанный на правилах*, заключается в поиске с использованием правил и шаблонов, которые составляются вручную. Эксперты составляют описания типов информации, которые необходимо извлечь. Подход хорош тем, что, при обнаружении ошибки, очень просто найти ее причину и внести необходимые изменения в правила. Проще всего составляются правила для унифицированных и стандартизированных объектов: имен, дат, наименований компаний и т.п.

Подход по правилам использует: применение лингвистических правил и шаблонов, содержащих лексическую и грамматическую информацию об извлекаемой конструкциях.

К недостаткам данного подхода можно отнести: трудоемкость создания правил, необходимость привлечения специалиста в области лингвистики, невозможность предусмотреть все варианты частных случаев, которые необходимо отразить в правилах.

Но вместе с тем есть и преимущества: уже созданные правила легко модифицировать и пополнять, в настоящее время правила зачастую описываются формальным языком, это позволяет легко в них ориентироваться.

Системы извлечения информации, использующие подход на правилах, опираются в своей работе на лингвистические шаблоны и правила, а также словарные ресурсы.

**Анализ существующих подходов к извлечению утверждений**

*Лингвистический шаблон* представляет собой формальное описание (образец) языковой конструкции, которую необходимо найти в тексте, чтобы извлечь нужную информацию. Например, N «работает в» NP, N «купил» N, где N – существительное, а NP (Noun Phrase) – группа существительного.

*Лингвистические правила* обычно состоят из двух частей. Левая часть правила (часть ЕСЛИ) содержит шаблон (образец) искомой языковой конструкции, а правая (часть ТО), в свою очередь, описывает действия, которые необходимо совершить, например, извлечь составляющие ее элементы (слова), приписать им определённую категорию и т.п.

*Онтологический подход* предполагает использование онтологии для извлечения отношений. По известному определению Т. Грубера, онтология – это спецификация концептуализации предметной области [5]. Концептуализация – это структура реальности, рассматриваемая независимо от словаря предметной области и конкретной ситуации.

Онтологии применяются сейчас довольно часто, и сфера их применения расширяется очень быстро. Одно из задач, которая решается с использованием онтологий является – организация поиска по смыслу в текстовой информации [6]. Текстовая информация пока еще остается основой документооборота. Объем этой информации очень велик, а задачи поиска и систематизации сложные и выполняют ответственную задачу.

Новыми задачами, связанными с извлечением знаний из текста, являются:

* формирование сообщений на заданную тему;
* извлечение новых фактов по интересующей теме;
* реализация виртуального собеседника.

Для извлечения фактов по интересующей теме целесообразно использовать онтологию ПрО.

В настоящее время распространено использование лингвистических онтологий, которые создаются на базе существующей в языке лексики и используются для построения модулей лингвистических процессоров.

Онтологический подход можно рассматривают как один из этапов синтаксического анализа. Он может применяться как совместно с правилами и шаблонами, так и самостоятельно. Отличие такого подхода заключается в ориентации используемых лингвистических описаний на конкретные предметные области.

К недостаткам можно отнести: относительная сложность интерпретации извлекаемых из онтологии данных, необходимость привлекать экспертов для создания онтологий, необходимость решения вопроса автоматического пополнения онтологии.

Плюсы использования онтологий. Онтологии представляют наиболее полное описание предметной области, а значит обеспечивают извлечение именно тех понятий и отношений, которые в ней представлены. Благодаря тому, что все данные в онтологии представляют собой связанный граф, мы можем пройти по нему и извлечь больше информации, чем при подходе на правилах. Можно использовать несколько онтологий, которые будут выполнять разные задачи.

Можно сделать вывод, что машинный подход более удобен для автоматической обработки текстов, но для русского языка не представлены достаточно хорошие средства составления размеченных корпусов знаний, поэтому удобнее использовать подход на правилах. Также стоит брать во внимание вид извлекаемой информации, для извлечения утверждений наиболее подходящими являются подходы на правилах, так как с помощью лингвистических правил, шаблонов и онтологий описать конкретную ПрО проще, чем в ручную размечать текстовые корпусы. Использованием комбинации лингвистических шаблонов и онтологий позволяет добиться достаточно высокой точности и полноты текста, так как шаблоны и правила позволяют описать стандартные варианты для заданной ПрО. А онтология расширяет и дополняет эти средства.

В статье были рассмотрены основные виды извлекаемой информации: сущности, отношения, факты и события. Представлен анализ подходов к извлечению информации из неформализованных текстов. Представлен обзор лингвистических средств, использующихся при подходе на правилах. Исходя из плюсов и минусов представленных подходов и средств был сделан вывод о целесообразности использования онтологий для повышения точности и полноты извлечения утверждений в дополнении к лингвистическим правилам и шаблонам.

**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.