Филологические науки УДК 81'32

ИЗВЛЕЧЕНИЕ И ИДЕНТИФИКАЦИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ В РУССКОМ ЯЗЫКЕ*

Брыкина Мария Михайловна,
к.ф.н., лингвист,
ЗАО «Эвентос»,
143005, Московская область, Одинцовский район,
г. Одинцово, ул. Вокзальная, д. 4
тел.: +7 (495) 987 21 82
научный сотрудник
Институт мировой культуры МГУ,
119991, Москва, ГСП-1, Ленинские горы,
МГУ, 1-й учебный корпус, к. 854
e-mail: m.brykina@gmail.com
Файнвейц Александра Вадимовна,
лингвист,
ЗАО «Эвентос»,
143005, Московская область, Одинцовский район,
г. Одинцово, ул. Вокзальная, д. 4
тел.: +7 (495) 987 21 82
студент,
Freie Universität Berlin, Берлин, Германия
e-mail: fainalex@yandex.ru
Толдова Светлана Юрьевна,
к. филол.н., научный сотрудник
Институт мировой культуры МГУ,
119991, Москва, ГСП-1, Ленинские горы,
МГУ, 1-й учебный корпус, к. 854,
тел. (495) 939-51-19,
старший научный сотрудник,
Научно-образовательный центр семантических технологий

* Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации в рамках государственного контракта № 07.524.11.4005 от «20» октября 2011 г., заключенного между Министерством образования и науки Российской Федерации и ЗАО «Эвентос»

НИУ ВШЭ,
e-mail: toldova@yandex.ru

Аннотация. *Настоящая статья посвящена базовым параметрам системы извлечения из текста именованных сущностей, основанной на словарях. Компонент извлечения именованных сущностей используется во многих приложениях, в частности, весьма перспективным направлением является пополнение данных семантического веба (например, LOD-онтологий) информацией из неструктурированных источников (текстов). Основным объектом нашего внимания являются методы разрешения различной омонимии для именованных сущностей, основанные на словарях и регулируемые эвристическими правилами. Такого рода система позволяет, во-первых, обеспечить достаточно высокую точность выделения объектов. Во-вторых, она дает возможность пользователю-неспециалисту модифицировать и обновлять предметную область. В-третьих, вновь вводимые объекты также могут выделяться с высокой точностью. В статье прдеставлена общая структура словарей, а также специфические для различных классов свойства синонимов, контекстных слов, выражений и объектов, которые могут обеспечить разрешение омонимии.*

Ключевые слова: *извлечение именованных сущностей, онтологическая омонимия, извлечение сущностей на основе правил.*

1. Введение

Одним из перспективных направлений развития сети Интернет является развитие технологий связанных открытых данных, что включает формирование стандартизованных онтологических баз знаний. Такие технологии позволяют связывать разнородную информацию об объекте, представлять ее в унифицированном виде. Они облегчают пользователю навигацию по различным базам данных и другим источникам, в которых упоминается объект его интереса. Объектами интереса, как правило, являются именованные сущности, такие как персоны, географические объекты, организации, продукты и др. При этом достаточно много информации об объекте интереса может быть представлено в сети в неструктурированных текстах. Таким образом, с одной стороны существуют открытые ресурсы – онтологии, такие, например, как Freebase, DBPedia и др.,

связанные между собой, в которых отражена в структурированном виде информация о разных объектах реального мира и их отношениях. С другой стороны, современный контент, доступный в сети, характеризуется высоким уровнем мобильности: ни один ресурс не в состоянии отразить информацию о вновь возникающих в информационном поле объектах, о новых открытиях и разработках, о персонах и организациях, связанных с ними. Возникает необходимость достаточно оперативного извлечения информации об объектах, интересующих конкретного пользователя, актуальных для конкретной отрасли науки. Особенно востребованы такие функции в системах извлечения информации для мониторинга состояния дел в наиболее активно развивающихся областях, таких, например, как нанотехнологии. Более того, необходимо, чтобы информацию, извлеченную из текста, можно было интегрировать в уже существующую базу знаний. Иными словами, необходим механизм, позволяющий извлекать информацию об объектах реального мира, которые интересны конкретному пользователю, унифицировать ее, т.е. отождествлять с одним и тем же объектом, независимо от многообразия способов наименования этого объекта в различных текстах.

Распознавание именованных сущностей в системах такого типа предъявляет целый ряд требований:

- это высокая точность и полнота распознавания относительно ограниченного множества объектов, интересующих пользователя;
- возможность изменения списков и классов объектов в соответствии с пользовательскими интересами;
- отождествление объектов, выделенных из текста, и объектов из онтологии.

Возникает вопрос, на каких принципах должна строиться архитектура системы, отвечающей вышеперечисленным требованиям. В настоящей статье рассматривается опыт разработки такой системы, основанной на словарях.

В нашем исследовании мы предлагаем систему атрибутов и правил, реализованную в рамках “ручного” подхода, т.е. систему с использованием словарей и онтологий. Основными достоинствами этой системы, на наш взгляд, являются:

- - предсказуемость поведения;
- - наличие у пользователя возможности обновлять словари в любой момент времени в любом объеме;
- - высокая точность.

Сложности, с которыми сталкивается любая система автоматической обработки текста, связана с языковой неоднозначностью на всех уровнях анализа. С одной стороны, для многих сущностей существует более одного наименования, с другой, - одно и то же наименование может использоваться для обозначения разных понятий и объектов реального мира. Как уже отмечалось выше, система ориентирована на возможность пополнения словарей неподготовленными пользователями. В связи с этим одной из основных задач при разработке методов разрешения неоднозначности являлась следующая: выделить минимальный набор простых контекстных признаков (как лексических, так и онтологических), поддающихся описанию в рамках инструкций для неподготовленного пользователя, которые бы позволили выделять пользовательские объекты с высокой точностью.

В предлагаемом исследовании мы последовательно рассматриваем разные случаи омонимии для основных классов именованных сущностей, а также классы контекстных слов и выражений и других типов ограничений на контекст, которых достаточно для обеспечения точности распознавания объектов основных трех классов объектов – Организаций, Персон и Местоположений – выше 95%.

2. История вопроса

Задача распознавания именованных сущностей является одной из наиболее разработанных и в то же время актуальных задач в рамках такого направления автоматической обработки текста, как извлечение информации из текста. Английский термин Named Entity (именованная сущность) был впервые использован в связи с постановкой задачи в рамках конференции MUC (Message Understanding Conferences) в 1995 г. [Grishman and Sundheim 1996]. Таким образом, данная задача имеет достаточно давнюю историю. Наиболее активно разработки в этой области велись для английского языка. Также проводились соревнования систем по распознаванию персон, организаций и географических названий (PLO) и для ряда других языков, таких как японский, испанский, немецкий (подробный обзор можно найти в [Nadeau, Sekine 2007]). Разработки в области распознавания базовых именованных сущностей проводились также и на материале русского языка (см. [Киселев, Ермаков, Плешко 2004] и [Крейдлин 2006]). Некоторый реестр русскоязычных систем представлен на сайте http://pullenti.ru/CompetitorPage.aspx. Как отмечается в литературе по извлечению именованных сущностей (см., например, [Nadeau D. And Sekine S. 2007]), существует два базовых подхода к данной задаче: подход, основанный на правилах, и подход, основанный на машинном.

В последнее время статистический подход в решении данной задачи более популярен. Особенно это касается разработок не для русского языка (см., например, [Burnesku, Paska 2006] и [Gentile et al. 2010]). Однако статистический подход имеет целый ряд недостатков и подходит не для всех задач. Во-первых, для обучения подобных систем необходим большой заранее сформированный корпус текстов, предварительно размеченный вручную. Этот подготовительный этап требует существенного вложения человеческих и временных ресурсов. Во-вторых, после каждого (порой, даже

незначительного) расширения системы, необходимо переобучение системы. Такой подход позволяет достаточно быстро и относительно качественно решать задачу распознавания неизвестных ранее именованных сущностей, пополнять базу новыми именованными сущностями. Однако он не всегда подходит для обеспечения высокой точности распознавания объектов по спискам пользователя. Как было сказано выше, существует целый ряд задач, в которых требуется достаточно высокая точность идентификации объектов в соответствии с базой знаний пользователя, возможность оперативно расширять список распознаваемых экземпляров объектов, а также отдельных классов объектов, не теряя при этом точности и контролируемости системы. Такие задачи достаточно часто решаются с использованием правил (см., например, [Кузнецов 2012]). Ниже рассматриваются конкретные проблемы, которые возникают в связи с данной задачей, а также обсуждаются некоторые архитектурные особенности системы, обеспечивающей ее решение.

3. Проблемы и конкретные задачи в системах извлечения именованных сущностей

При выделении именованных сущностей возникают следующие задачи:

1) распознать в тексте цепочку, которая называет некоторую именованную сущность;
2) решить, к какому классу именованных сущностей данная цепочка относится;
3) решить, какой конкретно объект из текста данное наименование обозначает, либо это новый объект, ранее не упомянутый в тексте;
4) решить, какой конкретно объект из базы знаний данное наименование обозначает, либо это новый объект, отсутствующий в базе знаний.

При этом задача 1) может решаться как задача выделения любых названий именованных сущностей без опоры на онтологии или специальные словари. В работе нас будет интересовать другой вариант этой задачи: методы распознавания объектов, заранее заданных в онтологиях предметной области и в специальных словарях.

Существование предзаданных онтологий объектов, с одной стороны, существенно упрощает задачу, с другой, множество проблем, связанных с языковой синонимией и омонимией, остаются также актуальными.

Во-первых, для выделения всех упоминаний именованной сущности в тексте необходимо ввести в словарь все необходимые синонимы. Разные синонимы обладают разной «различительной силой»: в разной степени позволяют однозначно идентифицировать объект реального мира. Так, например, *Московский государственный университет имени М. В. Ломоносова* однозначно называет единственный объект, а сокращение *МГУ* может быть использовано для обозначения различных университетов в городе Москве (ср., например, *МГУ сервиса и туризма*). То есть для идентификации объекта недостаточно выделить некоторую цепочку, совпадающую с одним из синонимичных его названий, необходимо также разрешить омонимию.

Можно выделить три типа омонимии, релевантные для обсуждаемой задачи: языковая омонимия (например, *(ОАО) “Металлист”* vs. *металлист* ‘поклонник определенного музыкального направления’); омонимия классов (*(город) Владимир* vs. *(имя) Владимир*); онтологическая омонимия ((политик) *Сергей Иванов* vs. (ученый) *Сергей Иванов*).

С точки зрения подхода, обсуждаемого в работе, для того, чтобы обеспечить разрешение неоднозначности вышеперечисленных типов, необходимо ответить на следующие вопросы:

- какова должна быть общая архитектура системы, обеспечивающей удобное ведение пользовательских словарей, и разрешение омонимии на основе данных словарей;
- как должны быть устроены пользовательские словари; какими функциями они должны обладать;
- каковы должны быть правила заведения в словари информации об объектах тех или иных классов: как должны быть устроены различные типы синонимов для каждого онтологического класса и подкласса объектов;
- какие должны быть предусмотрены атрибуты для работы механизма разрешения неоднозначности; какие типы диагностических контекстов следует учитывать для определенных классов объектов и различных типах их синонимов.

Задача настоящей статьи: проверить на материале трех наиболее разработанных классов объектов (организации, персоны и географические наименования), насколько минимальный контекстный подход к разрешению омонимии (семантический и прагматический контекст) позволяет разрешать омонимию достаточно надежно.

4. Общая архитектура системы и принципы организации словарей

Разработанная система идентификации в тексте объектов из пользовательских онтологий основывается на процессоре OntosMiner (см. [Минор, Старостин 2007]). Она включает в себя следующие базовые компоненты:

- токенизатор;
- морфологический анализатор;

- (вспомогательное) выделение Персон по эвристическим правилам и Организаций по правилам и на основе статистических алгоритмов;
- минимизация «эвристических» Персон и Организаций;
- применение системных и пользовательских словарей синонимов и выделение соответствующих синонимам фрагментов текста;
- верификация выделенных синонимов с учетом приписанных им характеристик на основе правил;
- минимизация словарных Персон и Организаций;
- присваивание весов выделенным объектам;
- формирование ttl-карты документа.

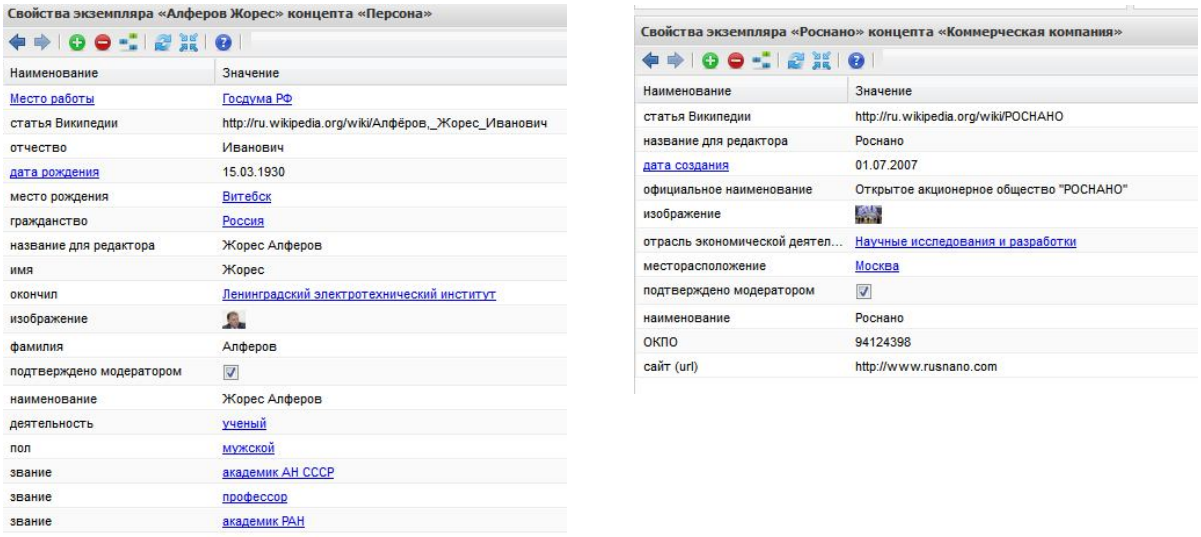
В данной работе основное внимание будет уделено подсистеме верификации выделенных синонимов.

Для того чтобы различать объекты внимания пользователя и их конкретные вхождения в текст мы пользуемся словарями двух типов:

- словари сущностей, которые мы хотим извлекать (далее – бизнес-объектов);
- словари поверхностного выражения объектов (словари синонимов для наименований объектов).

В бизнес-словарях (словарях сущностей) хранятся объекты внимания пользователя с необходимыми для них атрибутами, характеризующими их прежде всего как объекты внешнего мира (не лингвистически). В то же время часть этих атрибутов может быть необходима для дальнейшей работы системы. Эти словари могут отражать онтологическую структуру модели предметной области пользователя, например, информацию о том, что Персона работает или учится в некоторой Организации, которая, в свою очередь располагается в некотором Месте. На рис. 1 приведены образцы словарных карточек для Персоны и Организации. В значениях атрибутов

синим цветом с подчеркиванием выделены ссылки на другие объекты онтологии, как основные (*Ленинградский электротехнический институт; Москва*), так и вспомогательные (*ученый, профессор*).



Свойства экземпляра «Алферов Жорес» концепта «Персона»

Наименование	Значение
Место работы	Госдума РФ
статья Википедии	http://ru.wikipedia.org/wiki/Алфёров,_Жорес_Иванович
отчество	Иванович
дата рождения	15.03.1930
место рождения	Витебск
гражданство	Россия
название для редактора	Жорес Алферов
имя	Жорес
окончил	Ленинградский электротехнический институт
изображение	
фамилия	Алферов
подтверждено модератором	☑
наименование	Жорес Алферов
деятельность	ученый
пол	мужской
звание	академик АН СССР
звание	профессор
звание	академик РАН

Свойства экземпляра «Роснано» концепта «Коммерческая компания»

Наименование	Значение
статья Википедии	http://ru.wikipedia.org/wiki/РОСНАНО
название для редактора	Роснано
дата создания	01.07.2007
официальное наименование	Открытое акционерное общество "РОСНАНО"
изображение	
отрасль экономической деятел...	Научные исследования и разработки
месторасположение	Москва
подтверждено модератором	☑
наименование	Роснано
ОКПО	94124398
сайт (url)	http://www.rusnano.com

Рис. 1. Пример карточек бизнес-объектов Жорес Алферов, ОАО «Роснано».

Словари поверхностных выражений объектов включают синонимичные наименования одного и того же объекта. То есть словарными входами является подмножество синонимов для всех бизнес-объектов. Для каждого наименования (синонима) указывается ссылка на соответствующий бизнес объект. Так, например, для объекта ОАО “Роснано” синонимами являются:

- *РОСНАНО*;
- *RUSNANO*;
- *Российская корпорация нанотехнологий*.

Для каждого синонима при необходимости указывается лингвистическая информация, обеспечивающая выделение этого синонима в тексте во всех возможных формах: структура словосочетания и словоизменительная парадигма каждого элемента словосочетания.

Некоторые синонимы вне контекста могут однозначно указывать на соответствующие бизнес-объекты, т.е. взаимно-однозначное соответствие

можно провести без дополнительных проверок, в этом случае никакие дополнительные атрибуты этим словарным входам не нужны. Если же вне контекста значение синонима неочевидно (см. типы омонимии, перечисленные в п.3), то для правильной дальнейшей обработки текста необходимо указать, какая именно информация из текста может верифицировать данный синоним, то есть подтвердить, что в данном случае он указывает именно на тот бизнес-объект, к которому относится. В рассматриваемой системе это реализуются путем введения для синонимов специальных атрибутов, которые будут рассмотрены ниже.

При ведении словарей синонимов мы придерживались следующих принципов:

- в качестве входа в словаре синонимов вводится минимальный языковой материал, необходимый для различения именованных сущностей; в частности, для географических объектов не вводятся имена нарицательные – обозначения типа населенного пункта, например, для объекта ‘Московская область’ в словарь вводится только прилагательное *Московская* (подробнее см. п.5);
- кавычки не включаются в состав синонима; если они необходимы, это указывается в отдельном атрибуте, чтобы была возможность учесть все типы кавычек;
- регистр букв имеет значение, если только слово не записано целиком в нижнем регистре.

В реализованной системе содержится 5000 бизнес-объектов Местоположений (6500 синонимов), 3000 бизнес-объектов Персон (4000 синонимов), 5000 бизнес-объектов различных Организаций (10000 синонимов).

Ниже остановимся более подробно на отдельных классах объектов и рассмотрим, какова необходимая информация при ведении словарей данных

классов объектов. При этом основное внимание будет уделяться способам подтверждения омонимичных синонимов, хотя в системе предусмотрены и средства для отклонения тех синонимов, которые выступают в контексте, в котором не может находиться соответствующий класс бизнес-объектов. Например, именами людей часто называют улицы (*улица Сергея Эйзенштейна*), а местоположения могут фигурировать в наименованиях валют (*доллар США*).

5. Обработка географических объектов

5.1. Типы синонимов для географических объектов

Синонимами объектов административно-политической географии могут быть официальные, «бытовые» названия, аббревиатуры: *Московская область, Подмосковье, МО*. Согласно принципу подбора минимального языкового материалы во входы для географических названий не включаются слова, обозначающие тип географического объекта (*область, район, город...*). Словарным входом для синонима *Московская область* будет прилагательное *Московский* в форме женского рода единственного числа. Именно в такой форме данный синоним встречается в сочинительной конструкции: *в Московской и Ивановской областях*. В результате, такой подход позволяет обходить проблему распознавания географических названий в сочинительных конструкциях. Безусловно, одного прилагательного *Московская* не достаточно для однозначного распознавания географического объекта. Данный словарный вход должен обязательно проверяться на наличие или непосредственно рядом или в качестве вершины сочинительной конструкции слова, обозначающего соответствующий тип географического объекта.

5.2. Омонимия географических объектов

5.2.1. Омонимия двух Местоположений

Нередко названия районов, реже – городов или городов и стран, довольно часто – более мелких населенных пунктов бывают омонимичны. Так в России есть восемь Советских районов. Чтобы выделить в тексте правильный Советский район, необходимо учитывать информацию о вышестоящих административных объектах из онтологии. Если для синонима Местоположения указано, что необходимо подтвердить регион, то этот синоним подтвердится, только в случае успешной работы алгоритма по поиску в тексте его родительского или сестринского объекта.

Нередко встречается также омонимия двух географических объектов разных типов, например *город Тунис* и *страна Тунис*. В этом случае для соответствующих лукапов указывается, что должен быть реализован алгоритм поиска в ближайшем контексте подходящего родового слова, обозначающего тип географического объекта. Существует вспомогательный пользовательский словарь, в котором хранятся наименования для типов географических объектов также со своими синонимами (например, для объекта «область» синонимами будут слова *область, обл.*).

5.2.2. Омонимия Местоположения и имени или фамилии Персоны

Различные местоположения нередко бывают омонимичны фамилиям или именам людей: *Антон Чехов* vs. *Чехов* (город); *Лион Измайлов* vs. *Лион* (город). Эта омонимия снимается средствами ресурса минимизации: если Местоположение вложено в Персону, Местоположение удаляется.

5.2.3. Омонимия Местоположения и нарицательного имени или другой именованной сущности

Местоположение может быть омонимично нарицательному имени (*город Находка – находка*) или может входить в состав наименования других именованных сущностей (*Мадагаскар – мультфильм «Мадагаскар»*). Как и в

случае омонимии двух Местоположений разных типов, для проверки того, что словарный вход такого типа действительно относится к географическому объекту, используется проверка на соседство синонима с подходящим для него родовым словом. Кроме того, не подтверждаются синонимы Местоположений, находящиеся в кавычках, так как почти любое Местоположение может стать названием какого-нибудь магазина, кафе и т.д.

6. Обработка организаций

6.1. Типы синонимов организаций

Мы рассматривали различные типы организаций: коммерческие, общественные, международные, научные, государственные и др. В качестве словарных входов для синонимов были использованы наиболее короткие варианты названий организаций (кириллицей или латиницей) без опоясывающих кавычек и аббревиатуры.

6.2. Омонимия организаций

6.2.1. Омонимия Организации и нарицательного имени или другой именованной сущности

Названия организаций могут быть омонимичны обозначениям других реалий окружающей действительности. Рассмотрим следующие примеры:

(1) «Три богатыря» оказывает транспортные услуги.

(2) Мувинговая компания Три Богатыря предлагает услуги по профессиональной перевозке грузов

(3) Три богатыря – известные былинные герои.

В (1) и (2) словосочетание *Три Богатыря* употреблено как имя собственное, а в примере (3) – как нарицательное. Несомненно, кавычки способствуют разрешению подобной омонимии: опоясывающие кавычки однозначно в таких случаях указывают на то, что «сомнительный» синоним – это название организации. Однако, такое средство различения не может помочь абсолютно во всех случаях: как видно из примера (2), кавычки часто опускаются. Другой более надежный критерий, по которому можно однозначно

вычислить названия организаций, – это левый контекст. Если слева от синонима стоит обозначение типа организации (*ОАО, завод, финансовый холдинг)* или другое вспомогательное понятие (*директор, бухгалтер, пресс-секретарь*), можно с высокой вероятность утверждать, что синоним – собственное имя. Возможные вспомогательные слова и словосочетания, помогающие однозначно разрешить омонимию такого типа, задаются закрытым списком, программно реализованным как вспомогательный словарь. В Таблица 1 представлена общая схема оформления словарного входа «Три богатыря». Наиболее оптимальным будет указать, что для однозначного определения, является ли этот синоним названием компании, нужно, чтобы либо он был в опоясывающих кавычках, либо ему предшествовало подходящий родовой термин.

Таблица 1. Общая схема оформления словарного входа компании «Три богатыря»

синоним	***Атрибуты***	***соответствующий бизнес-объект***
Три богатыря	*- нужны кавычки;* *- слева нужно слово или словосочетание из закрытого списка для организаций*	Мувинговая компания «Три Богатыря»

6.2.2. Омонимия двух Организаций

Различные организации, деятельность которых разворачивается в разных отраслях экономики, могут иметь одинаковые названия. В таком случае для любого вхождения подобного «универсального» названия мы должны понимать, какая именно организация имеется в виду. К сожалению, опоясывающие кавычки или общие вспомогательные левоконтекстные понятия такие, как *главный руководитель, компания, продукция*, указывающие лишь на то, что синоним – имя собственное, не помогут различить подобную омонимию. В случаях, подобных (4) и (5), самую важную роль играет отрасль экономической деятельности, в которой работает компания и к которой должно относиться левоконтекстное слово

или словосочетание (см. Таблица 2). Каждой экономической отрасли соответствует закрытый список понятий, который может дополняться или сокращаться пользователем. Это могут быть относительные прилагательные, образованные от названий отраслей, например, *металлургический, банковский, строительный* и т.п.

(4) ХК "Северсталь" возглавляет турнирную таблицу.

(5) Горнодобывающая компания "Северсталь" владеет Череповецким металлургическим комбинатом.

Таблица 2. Контексты, разрешающие онтологическую омонимию: хоккейный клуб «Северсать» vs. компания «Северстаь». Фрагменты карточек объектов

Синоним	***Атрибуты***	***Соответствующий бизнес-объект***
Северсталь	- слева нужно слово или словосочетание из закрытого списка для спортивной отрасли	Хоккейный клуб «Северсталь»
Северсталь	- слева нужно слово или словосочетание из закрытого списка для горнодобывающей отрасли	Сталелитейная и горнодобывающая компания «Северсталь»

Таким образом, каждый вход вспомогательного словаря снабжается атрибутом, значением которого либо является отрасль, к которой относится данное понятие, либо метка о том, что однозначно отрасль определить нельзя (Таблица 3).

Таблица 3. Фрагмент вспомогательного словаря. Ключевые слова для отрасли

бизнес-объект ключевых слов для организаций	***отрасль***
хоккейный клуб	спорт
кондитерская компания	пищевая промышленность
Завод	неоднозначна

Государственные структуры, расположенные в населенных пунктах, но выполняющие схожие функции, тоже могут иметь одинаковые названия. В таких случаях предлагается для каждого из синонимов указывать соответствующий регион. Для установления соответствия между синонимом и бизнес-объектом необходима проверка совпадения локации синонимичного словарного входа и локации, встречающейся в документе наиболее близко к синониму. Так, в примере (6) должен определяться бизнес-объект «Министерство иностранных дел Российской Федерации», а в примере (7) – «Министерство иностранных дел Франции». Если в русскоязычном тексте указание на регион отсутствует, можно определить дефолтное значение и в примерах типа (8) проводить соответствие с «Министерством иностранных дел Российской Федерации».

(6) Министерство иностранных дел РФ ответило на просьбу Геннадия Онищенко.

(7) Владимир Путин посетил Париж и провел встречу с главой МИДа.

(8) Министерство иностранных дел ответило на просьбу Геннадия Онищенко.

Даже если для целей нашей работы релевантно распознать министерство иностранных дел только одной страны, необходимо учитывать то, что существуют министерства иностранных дел других стран для того, чтобы можно было учитывать потенциальную онтологическую омонимию.

6.2.3. Возможный вариант структуры вспомогательных словарей

Для более эффективной обработки текстовых документов, а также для экономии усилий мы предлагаем структурированный вариант оформления входов во вспомогательных словарях. К атрибуту «отрасль» можно добавить атрибут «тип входа», который может принимать одно из нижеперечисленных значений:

- префикс (*ЗАО, финансовая компания, фабрика)* обозначает, что последующий за ним синоним является названием организации;

Общий префикс не дает возможности определить конкретную экономическую отрасль;

- ключевое слово (*директор, управляющий*) обозначает, что следующий за ним синоним весьма вероятно является названием организации;
- ключевое прилагательное (*космический, горнодобывающий*) указывает на то, что атрибут «отрасль» следующего за ним общего префикса принимает конкретноотраслевое значение;
- постфикс (*Лимитед, & Со*) обозначает, что находящийся справа синоним является названием организации.

7. Обработка Персон

7.1. Типы синонимов для Персон

В русских новостных текстах для обозначения Персон обычно используется имя (иногда также отчество) и фамилия или же просто фамилия. Исходя из принципа подбора минимального языкового материала для синонима, в качестве синонима в таких случаях используется только фамилия. При этом, естественно, необходимо проверять такие синонимы на наличие в пре- или постпозиции подходящего соответствующему бизнес-объекту имени. Это может быть непосредственно имя, сочетание имени и отчества, инициалы. По дефолту фамилия без имени идентифицируется с тем же объектом, с которым идентифицируется эта же фамилия в сочетании с именем. Более сложный алгоритм нужен, только если в тексте упоминается несколько Персон, носящих одну фамилию (например, *Мишель Обама* и *Барак Обама*). Исключение из описанного процесса составляют широко известные Персоны, для обозначения которых могут использоваться исключительно фамилии. В таких случаях можно использовать алгоритм «непротиворечивости имени»: если в тексте нет сочетания имени и фамилии-

синонима, в котором имя не соответствует имени бизнес-объекта, то считать данную фамилию верифицированным синонимом для данной Персоны.

Кроме фамилий, в качестве синонимов (которым не нужна проверка на корректность имени) могут использоваться:

- псевдоним, сценическое имя или прозвище человека (*Борис Акунин, Витас*);
- сочетание имени и фамилии – это актуально, когда существует несколько варианта правописания имени или когда у человека несколько имен и могут использоваться разные их сочетание (*Абдалла ибн Абдель Азиз Ал Сауд, Абдуллах бин Абдул Азиз ал Сауд, Абдель Азиз Аль-Сауд, ...);*
- имя человека и его статус (*Королева Елизавета, Патриарх Московский и Всея Руси Кирилл*);
- имя иностранной Персоны, написанное латиницей (*Britney Spears*).

7.2. Омонимия Персон

7.2.1. Омонимия имени или фамилии Персоны и Местоположения

Случаи омонимии имени и фамилии рассматриваются в п. 5.2.

7.2.2. Омонимия имени и фамилии двух Персон

Омонимичные персоны не так уж часто встречаются в пределах одной предметной области, однако, система, безусловно, должна предусматривать возможность различения таких объектов. При этом используется один из следующих онтологических типов информации, которые должны быть приписаны бизнес-объекту:

- профессия, академическое или военное звание персоны, титул (постоянный статус Персоны);
- место работы (Организация или географический регион (для лидеров стран, областей));

- должность;
- (разные предметные области могут иметь свои специфические типы информации, которые могут быть полезные для идентификации Персон, например, для ученых релевантной может быть область их исследований).

Неоднозначный синоним проверяется на наличие в пределах предложения (или абзаца) основных или вспомогательных объектов перечисленных классов, которые подтверждают, что он в данном контексте обозначает именно данный бизнес-объект. При этом для должностей и статусов Персон необходимо вести свои словари с бизнес-входами и синонимами. Так, для бизнес-входа «профессор» указываются синонимы *профессор, проф.*

В пределах проведенного обследования мы не наблюдали других типов системной омонимии, которые бы требовали дополнительных источников информации для выделения или идентификации Персон.

8. Универсальный шаблон атрибутов синонимов для новых объектов

В процессе разработки предметной области может встать вопрос о необходимости выделения новых классов сущностей: например, различные спортивные команды или формы химических соединений. Весьма вероятно, что среди них окажутся объекты, название которых совпадает с названием других целевых объектов или других объектов окружающей действительности.

Предположим, вышеперечисленная система контекстов недостаточна для однозначного определения онтологического класса объекта, названного некоторым синонимом. Иными словами у нас есть неподтвержденный синоним, который не позволяет провести взаимно-однозначное соответствие к определенному бизнес-объекту. Например, у нас появился новый класс объектов – наноструктуры и синоним *наноструктура* для класса трехмерных

наноструктур. Контекстом, определяющим класс объекта, будет прилагательное *трехмерная*. Если на данный момент такой тип контекстов не предусмотрен для разрешения омонимии, пользователь может воспользоваться дополнительной возможностью: завести дополнительные условия проверки. Это можно сделать с помощью системы универсальных атрибутов.

8.1. Универсальный атрибут «Требуется объект»

В атрибуте прописывается объект, наличие которого во фрагменте текста подтверждает данный синоним. Объект может соответствовать любому бизнес-объекту в уже созданном фрагменте онтологии: основному (Организации, Персоне и т.д.) и вспомогательному (отрасли производства и т.д.). Программная реализация атрибута – ссылка на необходимый объект. Если в указанном фрагменте текста объект не будет найден, синоним останется неподтвержденным.

8.2. Универсальный атрибут «Требуется подстрока»

В атрибуте прописывается фрагмент текста, наличие которого подтверждает данный синоним. Если искомая подстрока не найдена, синоним считается неподтвержденным, бизнес-объект не выделяется.

8.3. Универсальный атрибут «Запрещается объект»

Аналогичен атрибуту *Требуется объект*. Отличие заключается в том, что в данном случае в значении атрибута прописывается ссылка на бизнес-объект, наличие которого во фрагменте текста запрещает подтверждение данного синонима. Если в тексте встречается этот объект, рассматриваемый синоним остается неподтвержденным.

8.4. Универсальный атрибут «Запрещается подстрока»

Аналогичен атрибуту *Требуется подстрока*. В атрибуте прописывается фрагмент текста, наличие которого запрещает подтверждение данного

синонима. Если искомая подстрока найдена, синоним считается неподтвержденным.

В каждом из вышеперечисленных атрибутов можно добавить вспомогательную функцию выбора сферы его действия, например, предложения, абзаца или всего документа. Сфера действия обозначит, на каком именно отрезке текста необходимо искать соответствующий объект или соответствующую подстроку.

8.5. Логические связки между атрибутами

Важное свойство универсальных атрибутов – программно реализованная возможность комбинирования их друг с другом и между собой:

- синоним подтверждается только при одновременном срабатывании всех перечисленных атрибутов (логическое И);
- для подтверждения синонима достаточно одного сработавшего атрибута (логическое ИЛИ).

Например, в рамках работы над новым концептом «Математические объекты» может потребоваться разграничить два понятия: «геометрический» вектор в n-мерном пространстве и «алгебраический» собственный вектор. Для этого необходимо завести два бизнес-объекта («собственный вектор» и «пространственный вектор») и два соответствующих им синонимичных входа (*вектор*), а затем разграничить их с помощью универсальных атрибутов так, чтобы в предложении (9) *вектор* отсылал к сущности «собственный вектор», а не к сущности «пространственный вектор».

(9) Ненулевой вектор s является собственным для преобразования A, если его образ A(s) колинеарен прообразу s.

Данная проблема может быть решена следующим образом: при входе, соответствующем бизнес-объекту «собственный вектор», мы указываем обязательность подстроки *собственн-* в предложении, а при входе,

соответствующем бизнес-объекту «пространственный вектор», – обязательность отсутствия подстроки *собственн-* (см. Таблица 4).

Таблица 4. Примеры карточек синонимов *вектор*

словарный вход	***соответствующий бизнес-объект***	***Атрибут***	***сфера действия атрибута***
Вектор	«собственный вектор»	требуется подстрока *собственн-*	Предложение
Вектор	«пространственный вектор»	запрещается подстрока *собственн-*	Предложение

Омонимия между двумя типами векторов может быть разрешена и другими способами. Например, с помощью указания на то, что для однозначного опознания концепта «собственный вектор» в том же документе должен быть упомянут концепт «матрица». В этом случае необходимо также указать, как именно должны сочетаться атрибуты «требуется подстрока» и «требуется объект» друг с другом: с помощью логического И (и тогда объект «собственный вектор» будет считаться подтвержденным тогда и только тогда, когда будут найдены и подстрока *собственн-*, и объект «матрица») или с помощью логического ИЛИ (тогда достаточно будет нахождения или объекта, или подстроки). Для случая, при котором должны быть выполнены все условия, карточка объекта будет выглядеть так (см. Таблица 5):

Таблица 5. Пример карточки синонима с несколькими атрибутами.

словарный вход	***соответствующий бизнес-объект***	***Атрибут***	***сфера действия атрибута***	***сочетаемость атрибутов***
вектор	«собственный вектор»	требуется подстрока *собственн-*	предложение	И
		требуется объект «матрица»	документ	И

9. Тестирование

Для оценки работы предложенной схемы выделения Персон, Местоположений и Организаций по словарям мы вручную разметили 300 новостных текстов. Выбор такого жанра был не случаен: мы считаем, что в нем соблюдается необходимый баланс между т.н. свободным и научным стилями. В первом случае очень часто возможно нарушение пунктуационных и орфографических правил русского языка, а также употребление редких разговорных выражений и синтаксических конструкций. В научных текстах, как правило, наблюдается, наоборот, ограниченное количество синтаксических конструкций и определенный лексический набор, что обеспечивает документам подобного типа более четкую структуру.

Каждый из текстов был размечен дважды, двумя разными людьми. Различия в разметке были перепроверены. Результаты тестирования приведены в Таблица 6

Таблица 6. Результаты тестирования

тип объекта	*Общее число аннотаций*	*полнота*	*точность*	*F-мера*
Местоположение	2270	0,98	0,99	0,98
Организация	1654	0,93	0,95	0,94
Персона	453	0,94	0,99	0,96

Достаточно высокие результаты тестирования объясняются несколькими факторами. Во-первых, такой метод тестирования показывает некоторый средний уровень работы системы именно на тех объектах, которые хорошо описаны в словарях системы. Во-вторых, эти результаты связаны с тем, что в новостном потоке большой процент случаев упоминания объектов в тексте составляют «топовые» объекты. Поэтому достаточно очень точно учесть в системе случаи омонимии именно для данных объектов. При этом значительный процент объектов встречается в новостном потоке крайне редко. Следовательно, при том, что для менее информационно значимых объектов случаи омонимии более частотны, отсутствие соответствующих

проверочных контекстов для них слабо влияет на общее качество работы системы. Таким образом, если наиболее упоминаемые объекты описаны хорошо, разные случаи омонимии для их синонимов учтены, то можно добиться достаточно высокой точности и полноты с точки зрения общей оценки работы системы.

Необходимо отметить, что качество работы системы сильно зависит от качества предварительной работы со словарями синонимов: необходимо, чтобы были учтены все возможные синонимы для бизнес-объекта.

При необходимости система может быть протестирована иначе: может учитываться не общая оценка работы объекта в целом в произвольном наборе текстов, а метрика по каждому из бизнес-объектов в каждом из типов. Можно предложить два подхода к такому более точечному тестированию:

- «прямой»; в этом случае для каждого из бизнес-объектов отбирается определенное количество текстов X, в которых должно быть установлено наличие бизнес-объекта, а также Y текстов, в которых бизнес-объект не должен быть выделен; для каждого из бизнес-объектов могут быть подсчитаны точность, полнота и F-мера;
- «функциональный»; в этом случае всё множество бизнес-объектов в зависимости от набора атрибутов разбивается на подмножества, и тестируется только по одному представителю от каждого подмножества.

10. Результаты

В статье была представлена система распознавания именованных сущностей, основанная на словарях. В задачи исследования входило рассмотреть различные ситуации языковой и онтологической омонимии, возникающие при распознавании в текстах именованных сущностей разных классов (персон, организации, географические названия и др.). В данной работе мы обобщили минимальные правила проверки контекстных

ограничений для разрешения разных типов омонимии и рассмотрели опыт работы с системой, в которой реализована возможность неподготовленного пользователя работать со словарями и бизнес-онтологиями. Правила поддаются описанию в виде инструкции и позволяют пользователю самостоятельно вводить необходимые данные об интересующих его объектов для того, чтобы система смогла правильно их распознавать. В работе были рассмотрены возможные типы синонимов для различных классов объектов, а также роль следующих характеристик, обеспечивающих разрешение омонимии:

- текстовые маркеры (например, кавычки);
- онтологические свойства объектов;
- иерархия онтологических типов объектов;
- географические объекты, ассоциированные с данным объектом;
- атрибуты, специфичные для отдельных классов объектов (статус-роль для персон, отрасль для организаций и т.п.).

Параметрами проверки контекстных признаков может быть при этом ближайший контекст, предложение, абзац, целый текст.

Как показали наши эксперименты, возможно создать унифицированную систему контекстных признаков и правил ведения словарей синонимов и контекстных проверок, которая обеспечивает достаточно высокую точность работы системы относительно объектов, интересующих пользователя.

11. Благодарности

Мы хотели бы поблагодарить нашего программиста Петра Жалыбина, который является одним из инициаторов представляемого подхода и разработчиком программного обеспечения, наших коллег Анастасию Бонч-Осмоловскую, Андрея Идиатуллина, Евгения Федько и Юлю Зинову за поддержку и продуктивную совместную работу.

Литература

1. Bunescu R. and Paska M. Using Encyclopedic Knowledge for Named Entity Disambiguation // Proceedings of of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006). - Trento, Italy: Association for Computational Linguistics, 2006, pp. 9-16.
2. Gentile A. L. et al. Cultural Knowledge for Named Entity Disambiguation: A Graph-Based Semantic Relatedness Approach // Serdica Journal of Computing. - Sofia, Bulgaria: Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 2010. 4: Vol. 2. pp. 217-242.
3. Grishman R., Sundheim B. Message understanding conference-6: A brief history //Proceedings of COLING. – 1996. – Т. 96. – С. 466-471.
4. Nadeau D. and Sekine S. A survey of named entity recognition and classification [Journal] // Linguisticae Investigationes. - Amsterdam, Netherlands: John Benjamins Publishing Company, 2007. - 1: Vol. 30. pp. 3-26.
5. Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в т ексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2004. – Москва, Наука, 2004
6. Крейдлин Л. Г. Программа выделения русских индивидуализированных именных групп TagLite // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2006. – Звенигород 2006, с. 292–297.
7. Кузнецов И. Методики выявления объектов и связей, заданных в неявном виде, 2012, URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/%D0%9A%D1%83%D0%B7%D0%BD%D0%B5%D1%86%D0%BE%D0%B2_%D0%98_%D0%9F.pdf
8. Минор С., Старостин А. Онтос: Технология извлечения знаний из неструктурированных текстов и семантическое индексирование // Компьютерная лингвистика и интеллектуальные технологии («Диалог-2007»). Труды международной конференции. – Бекасово, 2007.