К вопросу о разрешении семантической омонимии топонимов в русскоязычных текстах

К.К. Боярский[1], Е.А. Каневский[2], Д.А. Буторина[1]

[1] Университет ИТМО, [2]Институт проблем региональной экономики РАН

Boyarin9@yandex.ru, eak300@mail.ru, daanareevna20@gmail.com

Аннотация

Одним из способов анализа документов на естественном языке является извлечение именованных сущностей, в частности, топонимов. Особенность данной задачи заключается в необходимости разрешать омонимию топонимов с другими словами языка. Областью исследования являются русскоязычные тексты, имеющие общеполитический, художественный и узкоспециальный характер.

В качестве инструмента семантико-синтаксического анализа использовался парсер SemSin. Для снятия омонимии топонимов проводился поиск сопутствующих слов двух типов: определяющих, называющих классы географических объектов (река, город), и маркерных, образующих достаточно частотные словосочетаниями с топонимами (излучина, предместье).

Составлены «микрословари» маркерных слов для топонимов классов рек, городов и гор. Показано, что их использование в ряде случаев позволяет однозначно определить, что данное слово является топонимом соответствующего класса, в том числе для слов, отсутствующих в словаре. Полученные результаты не зависят от конкретного парсера и словаря и могут быть использованы для повышения точности технологий TextMining.

Ключевые слова: топонимы, омонимия, анализ текста, парсер, маркерные слова

Библиографическая ссылка: Боярский К.К., Каневский Е.А., Буторина Д.А. К вопросу о разрешении семантической омонимии топонимов в русскоязычных текстах // Компьютерная лингвистика и вычислительные онтологии. Выпуск 3 (Труды XXII Международной объединенной научной конференции «Интернет и современное общество», IMS-2019, Санкт-Петербург, 19 – 22 июня 2019 г. Сборник научных трудов). — СПб: Университет ИТМО, 2019. С. 19 – 28. DOI: 10.17586/2541-9781-2019-3-19-28

Введение

Средства массовой информации каждый день производят огромное количество данных, в частности, новостей, которые описывают различные явления, происходящие в том или ином регионе земного шара. Достаточно часто в региональных новостях важную роль играют так называемые **именованные сущности** (Named Entity). Выделение именованных сущностей – одна из ключевых задач извлечения информации (структурированных данных) из неструктурированных или слабоструктурированных документов [1]. Ее суть – найти в тексте названия или идентификаторы объектов определенного типа. Впервые задача была сформулирована еще в 1996 году на Message Understanding Conference, где в качестве сущностей рассматривались: организации, места, люди и некоторые числовые выражения. Позднее она рассматривалась на конференциях Conference on Computational Natural Language Learning (CoNLL) CoNLL-2002 и CoNLL-2003 [2].

Под термином **именованная сущность** сегодня понимается объект определенного типа, имеющий имя, название или идентификатор. Какие типы (классы) выделяет система, определяется в рамках конкретной задачи. Для новостных текстов обычно это PERSON – личности, ORGANIZATION – организации, LOCATION – географические объекты и MISCELLANEOUS – разное (числовые выражения, даты, события, слоганы и т.д.). В ряде случаев количество классов бывает и больше, так известны исследования по 29-ти классам [3] или даже по 200-м классам [4], в число которых входят продукты, книги, события, животные, растения, аэропорты и т. д.

Выделение именованных сущностей может использоваться для классификации текстов, сентимент-анализа, выявления парафраз и др. Важность этих задач (а также сложность их решения) подтверждается тем, что регулярно проводятся конкурсы соответствующих программных решений [1, 5, 6].

Важную роль при решении этих задач играет правильное выделение из текстов названий географических объектов (топонимов), к которым может иметь отношение данное сообщение. Как отмечалось в [7], проблема извлечения топонимов является достаточно сложной. Из-за омонимии одно и то же слово может обозначать географический объект или сущность, никакого отношения к топонимам не имеющую.

Постановка задачи

В качестве средства анализа текстов использовался парсер SemSin [8], осуществляющий глубинный синтаксический анализ русскоязычных текстов. Он использует расширенный вариант словаря В.А. Тузова [9], объемом свыше 194 тыс. лексем (около 170 тыс. слов). Классификатор расширен до 1700 классов. Анализ текста осуществляется под управлением продукционных правил [10, 11]. В процессе синтаксического анализа предложения одновременно выполняются снятие грамматической и частеречной омонимии, сегментация предложения и построение синтаксического дерева зависимостей. Достаточно часто разрешается и семантическая омонимия.

Парсер допускает настройку семантики на определенную предметную область. Для этого достаточно провести предварительное обучение парсера: разобрать несколько текстов для выбранной предметной области и выявить ряд омонимичных лексем, искажающих семантику разбора. Эти лексемы вместе с их классами помещаются в специальный файл, что обеспечивает автоматическое исключение их при всех дальнейших сеансах анализа. Как показывает практика, количество такого рода лексем не очень велико: при настройке парсера на анализ текстов из достаточно специфической области исторического кораблестроения оказалось достаточно отобрать 35 лексем [12].

Результат работы парсера, представленный в виде xml-файла, содержит лемму и полную информацию о грамматических (часть речи, род, число, падеж…), синтаксических (типы связей) и семантических (класс по классификатору) характеристиках слов в предложениях. Эти данные представляют своего рода «полуфабрикат» и могут использоваться для дальнейшего анализа и извлечения любой интересующей информации. Поэтому важно не только повышать точность разбора, но и максимально уменьшить его неоднозначность.

Целью данной работы является исследование возможности определения следующих факторов:

— является ли слово, написанное с прописной буквы, топонимом?
— к какому классу топонимов относится данное слово?
— какая синтаксическая связь подключает данное слово к дереву подчинения?

При анализе топонимов необходимо иметь в виду, что никакой справочник не в состоянии охватить все географические названия. Поэтому надо быть готовым к тому, что в тексте встретится неизвестное название. Это может быть либо слово, вообще

отсутствующее в словаре, либо слово, имеющее другое значение. В данной работе мы не рассматриваем ситуацию, когда топоним совпадает со словом общей лексики (гора Белуха — дельфин белуха, река Тигр — животное тигр). В этом случае обычно достаточно произвести простейший графематический анализ по регистру слова.

Также не ставится задача определить географическое положение объекта. По-видимому, без детального анализа больших фрагментов текста эта задача в общем случае не разрешима. Так гора Верблюд есть в Пятигорске и на Камчатке, а Черных речек в 30-километровой окрестности Санкт-Петербурга по меньшей мере три. В любом случае это задача не для парсера.

В используемом нами семантическом классификаторе около 90 классов содержат слова, являющиеся именами собственными, из них 30 классов выделены для обозначений топонимов. В табл. 1. приведены укрупненные классы, содержащие топонимы или имена собственные, омонимичные топонимам, и число лексем в этих классах.

Таблица 1. Семантические классы имен собственных

Семантические классы	Число лексем
Сухопутные природные объекты (горы, пустыни, острова…)	700
Водные природные объекты (реки, моря, заливы…)	1200
«Города» — населенные пункты (города, поселки, улицы…)	4600
«Уезды» — административные единицы (страны, области, провинции, штаты…)	800
Астрономические объекты (звезды, созвездия, планеты)	140
Фамилии	13700
Имена	4400
Мифологические существа	120
Учреждения (производственные, научные, спортивные…),	2700
Техника (автомобили, корабли, оружие…)	300

При разборе текста делалась попытка по возможности однозначно определить, к какому классу относится сущность, обозначаемая словом с прописной буквы.

Для оценки качества разбора проводилось сравнение с результатами работы парсера ЭТАП-4 [13].

Семантическая омонимия

При анализе текста прежде всего производится предварительная обработка, в ходе которой текст делится на предложения и слова и определяются морфологические характеристике слов. Затем текст пословно подается на вход синтактико-семантического анализатора. Блок правил, обрабатывающий топонимы, подключается при обнаружении слов, начинающихся с прописной буквы. В дальнейшем будем называть это слово **целевым**.

К семантической омонимии мы относим ситуацию, когда целевое слово именует объекты разной природы: населенные пункты, реки, острова, планеты и т. д. Некоторые случаи семантической омонимии топонимов приведены в табл. 2. Отбор производился по Большому Энциклопедическому Словарю [14]. Совпадения названий сгруппированы по условным классам с детализацией несколько более подробной, чем в табл. 1.

В табл. 2 одной звездочкой отмечены случаи, когда в два класса попадает не более 10 совпадающих названий, двумя звездочками – от 11 до 20, тремя – более 20.

Заметим, что одно и то же слово может обозначать как топоним, так и «негеографическое» понятие (названий предприятий, имен и фамилий людей и т. д.), причем здесь графематический анализ ничего не дает.

Наиболее частотной являются ситуации, когда название населенного пункта совпадает с фамилией человека, таких случаев более ста. При этом город может быть назван в честь кого-то (*Курчатов*, *Корсаков*, *Пушкин*), или имеет место простое совпадение (*Лондон*, *Адлер*).

Совпадение город – имя носит обычно случайный характер (*Аделаида*, *Лида*, *Милан*).

Совпадения названий городов с названиями рек, как правило, связаны с географическим положением населенных пунктов (*Анадырь*, *Нарва*, *Сакраменто*).

Во многих случаях совпадают названия островов и расположенных на них островных государств (*Барбадос*, *Исландия*, *Куба*). То же относится и к ряду названий городов (*Диксон*, *Кадьяк*), хотя, например, г. Ванкувер не находится на о. Ванкувер.

Для определения семантики целевых слов производится анализ ближнего контекста.

Таблица 2. Омонимия топонимов

	Гора	Город	Залив	Звезда	Озеро	Остров	Планета	Регион	Река	Страна	Уезд	Фамилия	Имя	Учреждения	Другое
Гора								*	*		*		*	*	*
Город			*	*	*	**		*	***	**	**	***	**	**	**
Залив						*					*	*			*
Звезда										*		*	*	*	*
Озеро						*				*		*	*	*	*
Остров										**	*	*	*	*	*
Планета													*	*	*
Регион														*	*
Река										*	**	*	**	**	*
Страна												*	*	*	*
Уезд												*	*	*	*

Аббревиатуры

Прежде всего, проверяется наличие контактно слева аббревиатуры, которая может обозначать географический объект (г., о., оз., пос., р.), адрес (пер., пл., пр., просп., ул.) или личность (г., гг.). Если обнаруживается аббревиатура географического объекта, то слово с прописной буквы считается топонимом. Например, *р. Волга* будет именно рекой, а не маркой автомобиля или спортивным клубом.

Ситуация, как обычно, осложняется омонимией, на этот раз аббревиатуры. Аббревиатура «о.» достаточно легко дифференцируется от предлога (предлог без точки) и от инициала (инициал с прописной буквы). Остаются варианты «остров» или «отец». В парсере ЭТАП-4 во всех случаях выбирается «отец», даже в предложении *Я посетил о. Мадагаскар*. В нашем парсере «о.» трактуется как остров, целевое слово принимает класс островов независимо от наличия в словаре и его словарного значения. Например, *о. Суматра* (есть в словаре как остров), *о. Вознесения* (есть в словаре, но в другом значении), *о. Баратанг* (отсутствует в словаре). Остается проблема, как интерпретировать аббревиатуру в сочетаниях типа *о. Иван*. Оказывается, что островов, названных по мужским именам очень мало, так что разумно считать, что «о.» имеет значение «отец» только в том случае, если целевое слово является словарным именем мужского рода.

Еще более запутанной является ситуация с аббревиатурой «г.», что может означать «гора», «город», «господин», «год». ЭТАП-4 для целевых слов, имеющихся в словаре, дает интерпретацию «город», для несловарных – «год». В нашей системе при обнаружении слева от целевого слова аббревиатуры «г.» продолжается дальнейший анализ левого

контекста, и интерпретация «год» выбирается только при обнаружении цифрового токена. В остальных случаях интерпретация определяется классом целевого слова: в сочетании *г. Новгород* «г.» означает «город» и получает класс населенных пунктов, а в сочетании *г. Эверест* «г.» это уже «гора». Для слов, отсутствующих в словаре, оставляются все варианты. В некоторых случаях удается уточнить класс при дальнейшем разборе предложения.

Определяющие слова

Под определяющими словами будем понимать слова, непосредственно называющие класс топонима: *река, остров, город* и т. д.

Если есть определяющее слово, и топоним имеется в словаре, то снятие семантической омонимии не представляет трудностей. Так в сочетании *озеро Байкал* Байкал — озеро, а не напиток, связь «Название» (в ЭТАП-4 ей соответствует аппозитивная связь).

Если есть определяющее слово, но целевое слово отсутствует в словаре в качестве топонима (*Я приехал на остров Зуб*), то разбор правильный и в SemSin и в ЭТАП-4. Отметим, что в дереве подчинения, построенном в ЭТАП-4, отсутствуют семантические пометы, так что правильность семантического класса проверить нельзя. Если слово вообще отсутствует в словаре (*остров Аогасима*) — в SemSin правильная установка класса и типа связи, ЭТАП-4 дает связь «квазиагент», т.е. не определяет это слово как название.

Если топоним имеется в словаре, а определяющего слова нет, то снятие семантической омонимии сопряжено с большими трудностями, преодолеть которые удается только в отдельных случаях. Рассмотрим предложение *Я приехал на Лену*. После морфологического разбора словоформа *Лену* имеет четыре значения: дательный падеж от лексем *ЛЕН* (либо земельный надел, либо административная единица Швеции) и винительный падеж от лексем *ЛЕНА* (либо река, либо имя). Два первых варианта отбрасываются по результатам графематического анализа (они должны начинаться со строчной буквы), два последних остаются.

Однако в процессе построения полного дерева подчинения удается на основе анализа классов подключаемых актантов глагола «приехал» снять и эту омонимию и оставить только название реки. ЭТАП-4 дает для лексемы *ЛЕНА* помету «одушевленное». Разумеется, в предложении *Я увидел Лену* останутся оба варианта.

Маркерные слова

В некоторых случаях при отсутствии определяющего слова, можно попытаться установить семантику имени собственного по маркерным словам. Под маркерными словами мы будем понимать слова, часто употребляющиеся с топонимами определенного типа. Например, для рек такими словами могут быть *течение, русло, устье*, для гор — *вершина, склон*, для озер — *берег, остров* и т. д. Характерной особенностью маркерных слов является то, что они присоединяют к себе имя собственное по родительному падежу (в ЭТАП-4 это связь «квазиагент»), а не как название.

Нами был исследован вопрос, насколько точно то или иное маркерное слово позволяет однозначно идентифицировать семантический класс топонима. Результаты, полученные при экспертном анализе нескольких тысяч предложений, представлены в таблицах 3–5.

В качестве маркерных слов были отобраны слова из класса «Ландшафт», часто сочетающиеся с названиями рек, городов и гор. Затем была сделана выборка из Национального корпуса русского языка (НКРЯ) [13] с дополнительным условием, чтобы справа шло слово с прописной буквы (кандидат в топонимы). Дальше были исключены случаи, когда между маркером и кандидатом стоит какой-либо знак препинания. К полученному списку были добавлены предложения, извлеченные по тем же правилам из новостных, художественных и спортивных текстов общим объемом около 55 млн. слов.

Рассмотрим пример: *Он возглавил экспедицию в Маньчжурию с целью изыскания тропы, которая спрямляла бы излучину Амура*. Слово *Амур* как имя собственное имеет не менее пяти значений: река, планета, фирма, спортивный клуб, бог. С учетом маркерного слова *излучина* остается единственное значение — река.

Таблица 3. Маркерные слова для рек

Лексема	Всего в НКРЯ	Справа слово с прописной буквы	Отобрано для анализа	Процент сочетания с названием реки
Излучина	543	118	124	100%
Низовье	704	391	229	100%
Междуречье	211	81	95	99%
Приток	3534	941	480	99%
Верховье	1537	671	603	98%
Устье	6020	2242	743	97%
Дельта	905	291	243	90%
Пойма	2473	103	70	84%
Исток	3677	490	180	83%
Русло	4567	239	104	78%
Течение	51572	1225	224	52%
Водораздел	658	53	50	44%
Бассейн	7137	820	422	38%
Берег	75107	14275	770	23%
Вода	156992	3185	280	22%

Из табл. 3 видно, что одни маркеры с вероятностью 95% и больше сочетаются с названиями рек, а для других этот процент существенно ниже. К таким относится, например, слово *водораздел*. Это слово в русском языке имеет очень разнообразную семантику, и, как видно из таблицы, только в 44% случаев сочетается с названием реки. С другими топонимами и именами собственными этот маркер сочетается в 50% случаев: *…проходит водораздел Балтийского и Черного морей*, *…шли мы по водоразделу Кавказского хребта*, и даже *…поэтический водораздел Давида Самойлова с Борисом Слуцким*. Однако, учитывая, что практически все названия горных систем и морей есть в словаре, можно с уверенностью прогнозировать, что в предложении *…на водоразделе Мо и Сахэзы, местность принимает даже гористый характер* «Мо» и «Сахэзы» — названия рек. Это позволяет при разборе прогнозировать семантику отсутствующих в словаре слов с прописной буквы.

Таблица 4. Маркерные слова для городов

Лексема	Всего в НКРЯ	Справа слово с прописной буквы	Отобрано для анализа	Процент сочетания с названием города
Предместье			36	86%
Пригород			454	94%
Центр			476	41%
Только во множественном числе				
Кварталы	7349	180	224	74%
Бульвары	6845	74	46	56%
Улицы	95032	3599	498	55%
Вокзалы	4454	144	39	46%
Жители	30695	4607	520	46%
Парки	16005	144	176	43%
Переулки	10958	146	176	43%
Проспекты	6826	50	38	37%

Слово *берег* обычно сочетается с названиями рек, озер, островов, морей, стран, регионов. Поэтому в предложении *Лагерь разбили на берегу Яны* однозначно остается название реки, а не имя. В некоторых случаях уточнить разбор удается, расширяя контекст и анализируя прилагательные, стоящие слева от маркерного слова. Так удается отличить *минеральные в**о**ды* от *в**о**ды Волги*, и понять, что в предложении *Город вырос на обоих берегах Эльбы* речь идет о реке, а не об острове.

Следует отметить, что и здесь в некоторых случаях можно уточнить разбор, расширяя контекст и анализируя прилагательные, стоящие слева от маркерных слов *пригород* или *предместье*. Если перед этими словами стоит прилагательное, образованное от названия города, то целевое слово является собственно названием пригорода или предместья: *В парижском пригороде **Фонтенбло** состоялся аукцион личных вещей Наполеона* (название пригорода) vs *Он находится в пригороде **Парижа*** (связь по род. пад.). ЭТАП-4 в обоих случаях дает связь «квазиагент». В некоторых случаях удается даже снять семантическую омонимию топонима: *Ни дома, ни машины здесь в пригороде Ванкувера почти не запираются* (*Ванкувер* — город и остров, после анализа остается город).

Таблица 5. Маркерные слова для гор

Лексема	Всего в НКРЯ	Справа слово с прописной буквы	Отобрано для анализа	Процент сочетания с названием горы
Вершина	17801	1739	500	62%
Склон	8796	739	395	81%

Однако это родство не помешало манси спихнуть "старших братьев" с западных склонов Урала (*Урал* — река, горы, регион и спортивный клуб, после анализа остаются горы).

В некоторых случаях правильное определение топонима весьма затруднительно. Так в предложении *За ним в истоках Они и Талой начиналась его тайга* названия рек Она и Талая не выявляются ни нашим анализатором, ни ЭТАПом-4.

Заключение

Анализ употребления топонимов в русскоязычных текстах показывает, что степень их омонимичности достаточно высока. Автоматическое определение топонимов осложняется также наличием огромного количества географических названий, отсутствующих в словарях общей лексики.

Для разрешения этой задачи предложено кроме определяющих слов, непосредственно указывающих на класс топонима, использовать маркерные слова, косвенно позволяющие судить о наличии топонима. Так, маркеры *излучина*, *низовье*, *междуречье*, *приток*, *верховье* и *устье* в подавляющем большинстве случаев предшествуют названию реки. Их присутствие позволяет с точностью не менее 95% интерпретировать слово, отсутствующее в словаре, но начинающееся с прописной буквы, как название реки.

Отмечено, что даже определение слова с прописной буквы как топонима зачастую не позволяет однозначно определить тип географического объекта (табл. 2). Составлены «микрословари» маркерных слов для топонимов наиболее омонимичных классов: рек, городов и гор. Показано, что их использование позволяет повысить точность определения топонимов, а в некоторых случаях и установить их класс. Полученные результаты могут быть использованы для повышения эффективности технологий TextMining.

Литература

[1] Starostin A. S. et al. Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian / Starostin A. S., Bocharov V. V., Alexeeva S. V., Bodrova A. A., Chuchunkov A. S., Dzhumaev S. S., Efimenko I. V., Granovsky D. V., Khoroshevsky V. F., Krylova I. V., Nikolaeva M. A., Smurov I. M., Toldova S. Y. // FactRuEval 2016: Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). — М.: Изд-во РГГУ, 2016. С. 702-720.

[2] Tjong Kim Sang E. F., De Meulder F. Introduction to the conll-2003 shared task: language-independent named entity recognition. // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. V. 4, CONLL '03, Stroudsburg, PA, USA, 2003. P. 142–147.

[3] Brunstein A. Annotation Guidelines for Answer Types // BBN technologies (2002). URL: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.

[4] Shinyama Y, Sekine S. Named entity discovery using comparable news articles. // In Proceedings of the 20th international conference on Computational Linguistics, COLING '04, Stroudsburg, PA, USA, 2004. P. 848–853.

[5] Loukachevitch N. V., Rubtsova Y. V. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). Вып. 15 (22). — М.: Изд-во РГГУ, 2016. С. 416-426.

[6] Panchenko A. et al. RUSSE2018: a Shared Task on Word Sense Induction for the Russian Language / Panchenko A. , Lopukhina A. , Ustalov D. , Lopukhin K. , Arefyev N. , Leontyev A. , Loukachevitch N. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 30 мая — 2 июня 2018 г.). Вып. 17 (24), 2018. С. 546-564.

[7] Lieberman M.D., Samet H. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11. P. 1–36.

[8] Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SEMSIN // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15, №5. С. 869–876.

[9] Тузов В.А. Компьютерная семантика русского языка. СПб: Изд-во С.-Петерб. ун-та, 2004.

[10] Боярский К.К., Каневский Е.А. Язык правил для построения синтаксического дерева // Интернет и современное общество: Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». – СПб: ООО «МультиПроджектСистемСервис». 2011. С. 233–237.

[11] Боярский К.К., Каневский Е.А. Система продукционных правил для построения синтаксического дерева предложения. Прикладна лінгвістика та лінгвістичні технології: MegaLing-2011. К.: Довіра. 2012. С. 73-80.

[12] Artemova G. et al. Text Categorization for Generation of Historical Shipbuilding Ontology / G. Artemova, K. Boyarsky, D. Gouzйvitch, N. Gusarova, N. Dobrenko, E. Kanevsky, D. Petrova. // Communications in Computer and Information Science. 2015. Vol. 468. P. 1–14.

[13] Лингвистический процессор ЭТАП-4. URL: http://www. http://proling.iitp.ru/ru/etap4 (дата обращения: 22.02.2019).

[14] Большой энциклопедический словарь / Гл. ред. А.М. Прохоров. – 2-е изд., перераб. и доп. – М. : Большая Российская Энцикл., 1997.

[15] Национальный корпус русского языка. URL: http://www.ruscorpora.ru/ (дата обращения: 19.01.2019).

On the Issue of the Semantic Disambiguation of Toponyms in Russian Texts

K. Boyarsry [1], E. Kanevsky[2], D. Butorina [1]

[1] ITMO University, [2] Institute of Regional Economics Problems RAS,

One way of analyzing documents in natural language is to extract named entities, in particular, toponyms. The peculiarity of this task is the need to disambiguate of toponyms with other words of the language. The research area is Russian-language texts of general political, artistic, and highly specialized nature.
The SemSin parser was used as a tool for semantic-syntactic analysis. To remove the homonymy of toponyms, a search for related words of two types was carried out: defining, naming classes of geographical objects (river, city), and marker words that form quite frequent word combinations with toponyms (bend, suburb).
"Micro-dictionary" of marker words for toponyms of classes of rivers, cities and mountains were composed. It is shown that their use allows to define unambiguously that the given word is a toponym of the corresponding class, including for words that are absent in the dictionary. The results do not depend on the specific parser and dictionary and can be used to improve the accuracy of TextMining technologies

Keywords: toponyms, ambiguity, text analysis, parser, marker words

Reference for citation: Boyarsry K., Kanevsky E., Butorina D. On the Issue of the Semantic Disambiguation of Toponyms in Russian Texts // Computer Linguistics and Computing Ontologies. Vol. 3 (Proceedings of the XXII International Joint Scientific Conference «Internet and Modern Society», IMS-2019, St. Petersburg, June 19-22, 2019). - St. Petersburg: ITMO University, 2019. P. 19 – 28. DOI: 10.17586/2541-9781-2019-3-19-28

Reference

[1] Starostin A. S. et al. Evaluation of Named Entity Recognition and Fact Extraction Systems for Russian / Starostin A. S., Bocharov V. V., Alexeeva S. V., Bodrova A. A., Chuchunkov A. S., Dzhumaev S. S., Efimenko I. V., Granovsky D. V., Khoroshevsky V. F., Krylova I. V., Nikolaeva M. A., Smurov I. M., Toldova S. Y. // FactRuEval 2016: Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog». 2016. Vol. 15 (22).

[2] Tjong Kim Sang E.F., De Meulder F. Introduction to the conll-2003 shared task: language-independent named entity recognition. // In Proceedings of the seventh conference on Natural language learning at HLT-NAACL. 2003.Vol. 4. P. 142–147.

[3] Brunstein A. Annotation Guidelines for Answer Types // BBN technologies. 2002. URL: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.

[4] Shinyama Y., Sekine S. Named entity discovery using comparable news articles. // In Proceedings of the 20th international conference on Computational Linguistics, COLING ’04, Stroudsburg, PA, USA. 2004. P. 848–853.

[5] Loukachevitch N. V., Rubtsova Y. V. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis. // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog» (Moskva, 1–4 iyulya 2016 g.). M.: Izd-vo RGGU. 2016. Vol. 15 (22). P. 416-426.

[6] Panchenko A. et al. RUSSE2018: a Shared Task on Word Sense Induction for the Russian Language / Panchenko A., Lopukhina A., Ustalov D., Lopukhin K., Arefyev N., Leontyev A., Loukachevitch N. // Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog» (Moskva, 30 maya — 2 iyunya 2018 g.). 2018. Vol. 17 (24). P. 546-564.

[7] Michael D. Lieberman, Hanan Samet. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11. P. 1–36.
[8] Boyarsky K.K., Kanevsky E.A. Semantiko-sintaksicheskiy parser SEMSIN // Nauchno-tekhnicheskiy vestnik informatsionnyh tekhnologiy, mekhaniki i optiki. 2015. Vol. 15, №5. P. 869–876. (In Russian).
[9] Tuzov V.A. Komp'yuternaya semantika russkogo yazyka. SPb. Izd-vo S.-Peterb. un-ta, 2004. (In Russian).
[10] Boyarsky K.K., Kanevsky E.A. Yazyk pravil dlya postroeniya sintaksicheskogo dereva // Internet i sovremennoe obschestvo: Materialy XIV Vserossiyskoy ob"edinennoy konferentsii «Internet i sovremennoe obschestvo». – SPb.: OOO «Mul'tiProdzhektSistemServis», 2011. P. 233–237. (In Russian).
[11] Boyarsky K.K., Kanevsky E.A. Sistema produktsionnyh pravil dlya postroeniya sintaksicheskogo dereva predlozheniya. Prikladna lingvistika ta lingvistichni tekhnologii: MegaLing-2011. K.:Dovira, 2012. P. 73-80. (In Russian).
[12] Artemova G. et al. Text Categorization for Generation of Historical Shipbuilding Ontology / G. Artemova, K. Boyarsky, D. Gouzйvitch, N. Gusarova, N. Dobrenko, E. Kanevsky, D. Petrova. // Communications in Computer and Information Science. 2015. Vol. 468. P. 1–14.
[13] Lingvisticheskiy protsessor ETAP-4. URL: http://www. http://proling.iitp.ru/ru/etap4 (data obrascheniya: 22.02.2019). (In Russian).
[14] Bol'shoy entsiklopedicheskiy slovar' / Gl. red. A.M. Prohorov. – 2-e izd., pererab. i dop. – M. : Bol'shaya Rossiyskaya Entsikl., 1997. (In Russian).
[15] Natsional'nyy korpus russkogo yazyka. URL: http://www.ruscorpora.ru/ (data obrascheniya: 19.01.2019). (In Russian).