Н

УДК 004.89:004.93 DOI 10.24412/2413-7383-2024-4-16-24

А. В. Ниценко, В. Ю. Шелепов
Федеральное государственное бюджетное научное учреждение
«Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артёма, 118-б

ОБ ИСПОЛЬЗОВАНИИ СЕМАНТИЧЕСКОЙ ИНФОРМАЦИИ ДЛЯ СНЯТИЯ ОМОНИМИИ ИМЕНИТЕЛЬНОГО И ВИНИТЕЛЬНОГО ПАДЕЖА (КАК ЭЛЕМЕНТА СОЗДАНИЯ ОНТОЛОГИИ)

A. V. Nicenko, V. Ju. Shelepov
Federal State Budgetary Scientific Institution "Institute of Artificial Intelligence Problems", Donetsk
283048, Donetsk, Artyom str., 118-b

THE USE OF SEMANTIC INFORMATION TO DISAMBIGUATE THE NOMINATIVE/ACCUSATIVE HOMONYMS: AN ELEMENT OF CREATING ONTOLOGY

В статье предложен способ автоматического снятия омонимии именительного и винительного падежа существительных с использованием информации о семантических связях слов, извлеченной из большого корпуса текстов. Данные о семантических связях представлены в виде онтологии, состоящей из множества семантических триплетов или троек «субъект-предикат-объект». Результаты реализованы в экспериментальном программном обеспечении для снятия омонимии.
Ключевые слова: обработка естественного языка, снятие омонимии, онтология, граф знаний, семантическая тройка.

The article proposes a method for automatic disambiguation of nominative and accusative cases of nouns using information on semantic relationships between words extracted from a large corpus of texts. The data on semantic relationships are presented as an ontology consisting of a set of semantic triplets or triples "subject-predicate-object". The results are implemented in experimental software for disambiguation.
Keywords: natural language processing, disambiguation, ontology, knowledge graph, semantic triple.

Введение

Одной из ключевых задач в области обработки естественного языка является автоматическое снятие неоднозначности слов в текстах. Она заключается в выборе того значения многозначного слова, в котором оно употреблено в конкретном контексте. Неоднозначность, свойственная естественному языку, является серьёзным препятствием для компьютерного анализа текстов, поэтому разрешение многозначности широко используется в таких областях, как машинный перевод, автоматическое извлечение информации из текстов, информационный поиск и т.д.

В некоторых случаях снятие грамматической неоднозначности требует информации о семантических связях между словами. Аналогичные связи обычно закладываются в тезаурусы или онтологии. При снятии неоднозначности данная информация может использоваться для выбора семантически допустимого варианта в сомнительных случаях. Одним из таких видов неоднозначности является омонимия именительного и винительного падежа, которая напрямую влияет на установление роли субъекта и объекта в предложении.

Целью работы является разработка семантической базы и алгоритма снятия омонимии именительного и винительного падежа для двух существительных в сочетаниях с глаголом. Очевидно, что это важнейший момент создания онтологии для работы с текстами.

Связанные работы

Для снятия неоднозначности используются различные методы. Наибольшую распространенность получили вероятностные методы, которые учитывают статистические закономерности, выводимые из больших корпусов текстов с морфологической разметкой. Для русского языка такой метод представлен, например, в работе [1]. Одним из базовых методов является использование скрытых Марковских моделей (Hidden Markov Model, HMM). HMM анализирует последовательности слов и их тегов, выбирая наиболее вероятную последовательность. Эти методы часто называют N-граммными моделями, где биграммные модели анализируют последовательности из двух слов, а триграммные - из трёх [2]. Для английского языка эти методы работают достаточно хорошо и обычно демонстрируют не менее 96 % точности [3]. Для русского языка точность таких алгоритмов намного меньше, так как морфологическая омонимия в русском языке охватывает множество различных грамматических признаков.

С развитием и совершенствованием корпусов текстов с морфологической разметкой, таких как Национальный корпус русского языка и OpenCorpora, стали популярны методы статистического обучения. Стали появляться так называемые морфологические теггеры, основанные на модификациях HMM [4]. В работе [5] такие теггеры применяются для морфологического анализа русскоязычных текстов.

Метод условных случайных полей CRF (Conditional Random Fields) также используется для снятия морфологической многозначности. CRF представляет собой графовую модель, которая используется для представления совместных распределений набора нескольких случайных переменных. CRF, являясь разновидностью Марковских случайных полей, относится к дискриминативным вероятностным методам и не требует предположения независимости наблюдаемых переменных. В работе [6], например, CRF применяется для снятия морфологической многозначности, а в работе [7] показано, что качество морфологического анализа с помощью CRF превышает качество теггеров.

Последние годы для морфологического анализа стали применяться нейросетевые подходы [8], [9]. В одном из последних систематических обзоров [10] показано, что методы установления частей речи, основанные на нейросетевом глубоком обучении, превосходят по качеству все остальные. К недостаткам таких подходов относятся необходимость наличия больших обучающих выборок, сложность обучения и тонкой настройки, недетерминированность и неинтерпретируемость получаемых результатов [11]. Кроме того, для некоторых видов омонимии, как например омонимия именительного и винительного падежа существительных, их точность оказывается недостаточной [12].

Использование семантической информации для снятия омонимии

Омонимию именительного и винительного падежа имеют следующие категории существительных:

– неодушевленные существительные мужского рода в единственном и множественном числе (***стол*** *стоит* - *купили* ***стол***).

– неодушевленные существительные среднего рода в единственном и множественном числе (***окно*** *открылось – вижу* ***окно***), одушевленные в единственном числе (***животное*** *прячется – поймал* ***животное***).

– существительные женского рода 3 склонения, неодушевленные, в единственном и множественном числе (***дверь*** *закрылась – открыли* ***дверь***), одушевленные в единственном числе (***мышь*** *крадется – увидел* ***мышь***).

Именительный падеж существительного определяет его как подлежащее предложения – субъект в высказывании. Ситуация усложняется, если такое существительное является омонимом – кандидатом на именительный и винительный падеж. Как правило, выбор именительного падежа определяется в этом случае наличием согласованного глагола (сказуемого). Но возможна ситуация, когда существительных с такими свойствами два. Примеры: *«****огород*** *дает* ***урожай****», «****ледокол*** *прокладывает* ***путь****»*. Чтобы автоматически определить, где здесь именительный падеж (подлежащее, субъект), а где винительный (дополнение, объект), необходимо иметь дополнительную информацию о том, каким образом эти слова могут быть связаны семантически.

Смысловые отношения между словами можно описать в виде онтологии, состоящей из множества семантических триплетов или троек «субъект-предикат-объект». Субъект в нашем случае – это подлежащее, предикат – это глагол, который может связывать подлежащее и дополнение, а объект – это дополнение, связанное с субъектом через предикат (рис. 1). Таким образом семантическая база представляет собой, по сути, ориентированный граф знаний.

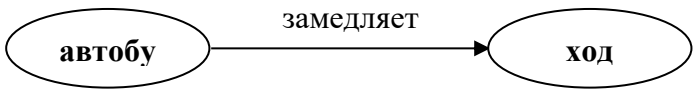

Рисунок 1 – Семантический триплет «Автобус замедляет ход»

Для получения смысловых отношений между словами мы использовали данные о совместном употреблении существительных и глаголов, извлеченные из Национального корпуса русского языка (НКРЯ) [13] со снятой омонимией. НКРЯ состоит из нескольких подкорпусов и представляет собой большую коллекцию текстов на русском языке общим объемом более 2 млрд. слов, оснащенную лингвистической разметкой и инструментами поиска. Каждый из подкорпусов является

большим по объёму и представительным, что делает их ценным материалом для количественных и качественных исследований. Основной подкорпус НКРЯ содержит около 30,5 млн. предложений и 374.5 млн. словоупотреблений, что составляет 17.9 % от общего объема корпуса. Из них со снятой омонимией – около 519 тыс. предложений и 6.1 млн. словоупотреблений (0.3 % от общего объема корпуса и 1,7 % объема основного подкорпуса) [14,15]. Из подкорпуса со снятой омонимией с помощью поиска было отобрано 270 тыс. предложений, содержащих два слова-омонима именительного и винительного падежа в сочетании с переходным глаголом (на основе разметки корпуса). Результат поисковой выдачи сохранялся в файл формата XLSX (рисунок 2), из которого программным способом извлекались данные в текстовом формате. При этом отбирались только те предложения, где данные слова не разделяются знаком препинания, и омонимы винительного падежа представлены без предлога [16]. Из отобранных предложений извлекались тройки: существительное в именительном падеже + глагол + существительное в винительном падеже (порядок и соседство не важны) и приводились к начальной форме (табл. 1).

H	R	S	T
Author	Ambiguity	Full context	
Алексей Полухин, Сергей Рябов	омонимия снят	Вчера пенсионная реформа вступила в новую фазу: **Минфин открыл** сезон приема заявок на **конкурс** по выбору управляющих компаний (УК) и	Минфин; открыл; конкурс;
Андрей Бахтин	омонимия снят	Однако **Минфин «заметил»** это только сейчас, подготовив **проект** бюджета с заниженным прогнозом курса евро	Минфин; заметил; проект;
Владислав Отрошенко	омонимия снят	Ее **облик вызывал** лишь полный и искренний **восторг**	облик; вызывал; восторг
Олег Павлов	омонимия снят	**Глаз ухватывал** в просвете только пройденный **туннельчик** -- теперь он задирался вверх и было видно, как тянулось к дневному свету поднято	Глаз; ухватывал; туннельчик;
Андрей Волос	омонимия снят	Излишне говорить, что каждый **просмотр повергал** Будяевых в смятение **и трепет**	просмотр; повергал; трепет
Людмила Улицкая	омонимия снят	В изученном им отрезке времени, от семнадцатого до пятьдесят шестого, в конкретном месте -- на территории СССР -- этот **фактор оказывал** с	фактор; оказывал; процесс
Сергей Викторов	омонимия снят	**Запад требовал**, чтобы Россия наращивала сырьевой **экспорт** и накапливала валютные ресурсы, иначе кредиторы отказывались реструктурирс	Запад; требовал; экспорт;
Анатолий Азольский	омонимия снят	Своего автотранспорта милиция никогда в избытке не имела, город по разнарядке ежедневно выделял ей машины, но сегодня в испуге от вче	переулок; запружен; Москвич;
Дуня Смирнова	омонимия снят	Первые два месяца западного вида **механизм выдавал** гражданам при входе в **центр** талончик с номером	механизм; выдавал; центр;
Ю. О. Домбровский	омонимия снят	Какой **ветер занес** вас в этот фруктовый **колхоз**	ветер; занес; колхоз
Юрий Трифонов	омонимия снят	Он посматривал сбоку на Динку **Абажур, видел** ее пунцовую щеку, вздернутый **нос**, черные кудри, выбившиеся из-под шерстяной лыжной ша	; Абажур; видел; нос;
Василий Шукшин	омонимия снят	По поводу чего **сбор**? -- **спросил** Лев Казимирыч, присаживаясь на **стул** к столу	сбор; спросил; стул;
Василий Аксенов	омонимия снят	**Привет**! -- **сказал** Алик и плюхнулся на **песок** рядом	Привет; сказал; песок;
Василий Гроссман	омонимия снят	Утром, едва она пришла на работу, ее вызвали к телефону, и чей-то любезный **голос попросил** ее зайти в паспортный **стол** городской милиции	голос; попросил; стол;
Василий Гроссман	омонимия снят	Какой хороший **вечер**, -- **сказал** Штрум, вдыхая сырой, холодный **воздух**	вечер; сказал; воздух
М. М. Пришвин	омонимия снят	К счастью для Травки, сильный **голод заставил** ее прекратить свой печальный **плач** или, может быть, призыв к себе нового человека	; голод; заставил; плач;
Ю. О. Домбровский	омонимия снят	Нет, не страх, никак не страх, никак не только один **страх заставил** меня перемахнуть в чужой **лагерь**	страх; заставил; лагерь
Ю. О. Домбровский	омонимия снят	**Раз**! -- **отсчитал** отец и загнул один **палец**	Раз; отсчитал; палец
Ю. О. Домбровский	омонимия снят	У него заколодо в боку, стали **жать** ботинки, воротник сделался узок и перехватил горло, кресло врезалось в тело, **пот покрыл** лицо, -- он накл	; пот; покрыл; таз
Ю. Н. Тынянов	омонимия снят	**Домик напоминал** более всего античный небольшой **храм**, но был устроен на крошечном острове среди озера, ранее бывшего болотом	Домик; напоминал; храм;
В. В. Вересаев	омонимия снят	**Поезд дал** свисток и начал замедлять **ход**	Поезд; дал; ход
И. С. Тургенев	омонимия снят	Уже не досада меня грызла, -- тайный **страх терзал** меня, и не один **страх** я чувствовал... нет, я чувствовал раскаяние, сожаление самое жгучее,	; страх; терзал; страх;
Д. В. Григорович	омонимия снят	**Народ, извещенный** случаем, валил на скотный **двор** со всех сторон и успел уже натискаться в избу вплоть до самых сеничек	Народ; извещенный; двор;
Владимир Василиненко	омонимия снят	Было тихо, что для наших мест редкость, довольно морозно, **градусник, прикрепленный** около небольшого оконца, показывал **минус** двадцат	градусник; прикрепленный; минус;
Владимир Василиненко	омонимия снят	Алюминиевый **чайник начал** выбрасывать из носика сначала **пар**, затем булькающий кипяток, когда появившийся на улице папа перехватил ин	чайник; начал; пар;
Наталья Емельянова	омонимия снят	Старинный **диван, обнаруженный** мной еще в прошлый **раз**, стоял прямо напротив печи	диван; обнаруженный; раз;
Михаил Карпач	омонимия снят	Районный **суд признал** объявление в розыск **и арест** Погодина незаконными, но с постановлением суда не согласилась районная прокуратура	суд; признал; арест;
Гузель Яхина	омонимия снят	**Мороз обжег** распаренные в теплой воде **лоб** и щеки, влажную еще кожу головы	Мороз; обжег; лоб;
Гузель Яхина	омонимия снят	С ликующим визгом устремлялась она теперь повсюду: под стол, где часто валялись упавшие листы с черновиками Баховых сказок (бумагой мс	песок; расчерченный; песок;
Гузель Яхина	омонимия снят	Ноздри при этом чуть пошевеливались -- казалось, фюрер ощупывает ими лакированное дерево: **нос оставлял** на узорчатом ясене влажный сл	нос; оставлял; след;

Рисунок 2 – Результат поисковой выдачи корпуса

Таблица 1 – Пример извлечения данных из корпуса

Текст из корпуса	Извлеченные данные
В эту самую минуту она замечает, что ***автобус (сущ. неод. ед. муж. им.)*** *чуть* ***замедлил (гл. перех.) ход (сущ. неод. ед. муж. вин.)***	***[автобус; замедлить; ход]***
Агентство (сущ. неод. ед. ср. им.) осуществляет (гл. перех.) *свою* ***деятельность (сущ. неод. ед. ср. вин.)*** *согласно Уставу агентства*	***[агентство; осуществлять; деятельность]***
Рассмотренный ***алгоритм (сущ. неод. ед. муж. им.) предполагает (гл. перех.)*** *параллельное* ***выполнение(сущ. неод. ед. ср. вин.)*** *всех его этапов и носит итерационный характер*	***[алгоритм; предполагать; выполнение]***

На основании этих данных была сформирована текстовая база, которая содержит семантические триплеты. Она состоит из записей вида:

автобус ход !
замедлить
замедлять
набирать
набрать

агентство деятельность !
заканчивать
осуществлять

алгоритм выполнение !
предполагать

Строка с восклицательным знаком содержит начальные формы существительных, первое из которых должно употребляться с перечисленными ниже глаголами в именительном падеже, а второе – в винительном. Необходимо отметить роль глагола в описанной базе. Ограничиться только парой существительных при решении вопроса, какое из них стоит в именительном падеже, а какое в винительном падеже невозможно. При изменении глагола падежи могут меняться местами, например*: Грузовик(им.) миновал шлагбаум(вин.). Шлагбаум(им.) пропустил грузовик(вин.).*

Дополнительно из частотного словаря [17] были отобраны 200 наиболее частотных существительных, имеющих полное совпадение форм в именительном и винительном падежах, из которых были образованы пары «субъект+объект» во всех возможных сочетаниях. Далее проводился поиск примеров совместного употребления полученных пар в корпусе текстов без снятой омонимии. Те пары слов, которые не встречаются в текстах вместе, исключались из списка. Также были исключены пары, которые вошли в полученную ранее базу из корпуса со снятой омонимией. Таким образом было получено около 5.5 тыс. новых пар. С этими парами была проделана большая работа по подбору подходящих по смыслу глаголов, с использованием НКРЯ и других источников информации. В результате было получено около 13 тыс. новых троек, которые пополнили текстовую базу.

Для программного снятия омонимии в рассматриваемой ситуации были разработаны следующие правила: пусть на отрезке текста существует такая тройка слов A,B,C (A и B существительные, C – глагол), что для них выполняются условия:

1) A[0].pos == «сущ» && A[0].case == «им» && A[1].pos == «сущ» & A[1].case == «вин»
2) B[0].pos == «сущ» && B[0].case == «им» && B[1].pos == «сущ» & B[1].case == «вин»
3) C[0].pos == «гл» && (C[0].verb_trans == «перех» || C[0].verb_trans== «пер/не»)

Здесь поле «pos» определяет часть речи, «case» - падеж, «verb_trans» - вид глагола (переходный или переходный/непереходный).

Тогда неоднозначность (A[0] | A[1]) и (B[0] | B[1]) снимается следующим образом:

если **согл_сущ_ гл**(A , C) == 1 && **согл_сущ_гл** (B , C) == 0, то (A[0] | A[1]) → A[0] , (B[0] | B[1]) → B[1]

если **согл_сущ_ гл**(A , C) == 0 && **согл_сущ_гл** (B , C) == 1, то (A[0] | A[1]) → A[1] , (B[0] | B[1]) → B[0]

В случае, когда в предложении оба существительных согласуются с глаголом, решение о выборе именительного и винительного падежа осуществляется на основании данных из семантической базы. Пусть на отрезке текста существует такая тройка слов A,B,C, что для них выполняются условия:

1) A[0].pos == «сущ» && A[0].case == «им» && A[1].pos == «сущ» & A[1].case == «вин»
2) B[0].pos == «сущ» && B[0].case == «им» && B[1].pos == «сущ» & B[1].case == «вин»
3) C[0].pos == «гл» && (C[0].verb_trans == «перех» || C[0].verb_trans== «пер/не»)
4) согл_сущ_ гл(A , C) == 1 && согл_сущ_гл (B , C) == 1

Тогда неоднозначность (A[0] | A[1]) и (B[0] | B[1]) снимается следующим образом:

если в базе существует тройка <A, C, B> то (A[0] | A[1]) → A[0] , (B[0] | B[1]) → B[1]

если в базе существует тройка <B, C, A> то (A[0] | A[1]) → A[1] , (B[0] | B[1]) → B[0]

Заключение

В статье представлен метод автоматического разрешения омонимии между именительным и винительным падежами существительных, основанный на использовании информации о семантических связях слов, извлеченных из большого текстового корпуса. Эти семантические связи представлены в виде онтологии, включающей множество триплетов «субъект-предикат-объект». Разработанные правила для снятия неоднозначности были реализованы на языке программирования C++ в экспериментальной программе. Всего на данный момент из корпуса текстов удалось извлечь около 122 тыс. семантических триплетов, куда входит около 4,7 тыс. переходных глаголов и 12 тыс. существительных. Учет семантической информации, заложенной в базе, позволил осуществлять снятие неоднозначности в случаях, когда без понимания смысловых отношений между словами сделать это невозможно.

Список литературы

1. Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог’2005. – 2005. – С. 188-197.
2. Сокирко А.В., Толдова С.Ю. Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп) // Интернет-математика 2005: автоматическая обработка веб-данных. – М., 2005. – С. 80-94.
3. Лакомкин Е.Д., Пузыревский И.В, Рыжова Д.А. Анализ статистических алгоритмов снятия морфологической омонимии в русском языке. [Электронный ресурс] URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf (дата обращения 20.12.2021).
4. Schmid H. Probabilistic part-ofispeech tagging using decision trees // New methods in language processing. 2013. Pp. 154-164.
5. Sharoff S., Kopotev M., Erjavec T., Feldman A., Divjak D. Designing and evaluating a Russian tagset // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). Vol. 26. Pp. 279-285.

6. Muzychka S.A., Romanenko A.A., Piontkovskaja I.I. Conditional Random Field for morphological disambiguation in Russian // Компьютерная лингвистика и интеллектуальные технологии. 2014. С. 455-465.
7. Антонова А., Соловьев А. Использование метода условных случайных полей для обработки текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии. – 2013. – С. 27–44.
8. Sorokin A., et al. MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian // Proceedings of the International Conference Dialogue 2017. 2017. V.1. P. 297-313.
9. Lyashevskaya O.N., et al. GRAMEVAL 2020 shared task: Russian full morphology and universal dependencies parsing // Proceedings of the International Conference Dialogue 2020. 2020. V.1. P. 553-569.
10. Chiche A., Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning aPproaches// Journal of Big Data. 2022. V.9. №1. P. 1-25.
11. Смирнов И.В. Интеллектуальный анализ текстов на основе методов разноуровневой обработки естественного языка. – М.: ФИЦ ИУ РАН, 2023. – 356 с.
12. Anastasyev, D. G. Part-of-speech tagging with rich language description / D. G. Anastasyev, A. I. Andrianov, E. M. Indenbom // Компьютерная лингвистика и интеллектуальные технологии : По материалам ежегодной Международной конференции "Диалог", Москва, 31 мая – 03 2017 года / Российский государственный гуманитарный университет. Vol. Выпуск 16 (23), Том 1. – Москва: Российский государственный гуманитарный университет, 2017. – P. 2-13.
13. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа: URL: https://ruscorpora.ru/ (дата обращения: 20.11.2024).
14. Копотев, М.В. Национальный корпус русского языка / М.В. Копотев, Л.А. Янда // Вопросы языкознания . – 2006. – № 5. – С. 149–155.
15. Жевнерович, Е.Э. Корпус текстов в научном исследовании / Е.Э. Жевнерович //Материалы II Международной научно-практической конференции «Лингвистика, лингводидактика, лингвокультурология: актуальные вопросы и перспективы развития» (Минск, 1–2 марта 2018 г.). – 2018. – С.25-32.
16. Ниценко, А. В. Использование данных Национального корпуса русского языка для снятия омонимии винительного и родительного падежа внутри парадигмы существительных [Текст] / А. В. Ниценко, В. Ю. Шелепов // Искусственный интеллект: теоретические аспекты, практическое применение: материалы Донецкого международного научного круглого стола. − Донецк: ФГБНУ «ИПИИ», 2023. – 252 с. – С. 137–141.
17. Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). [Текст] // М.: Азбуковник, 2009. –1112 с.

References

1. Zelenkov Yu.G., Segalovich I.V., Titov V.A. Probabilistic model of morphological homonymy removal based on normalizing substitutions and positions of neighboring words // Computer linguistics and intellectual technologies. Proceedings of the international seminar Dialogue. 2005. Vol. 2005. P. 188-197.
2. Sokirko A.V., Toldova S.Yu. Comparison of the efficiency of two methods for removing lexical and morphological ambiguity for the Russian language (hidden Markov model and syntactic analyzer of noun phrases) // Internet Mathematics 2005: automatic processing of web data. Moscow, 2005. P. 80-94.
3. Lakomkin E.D., Puzyrevsky I.V, Ryzhova D.A. Analysis of statistical algorithms for removing morphological homonymy in the Russian language. [Electronic resource] URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf (accessed 20.12.2021).4. Schmid H. Probabilistic part-ofispeech tagging using decision trees // New methods in language processing. 2013. P. 154-164.
5. Sharoff S., Kopotev M., Erjavec T., Feldman A., Divjak D. Designing and evaluating a Russian tagset // Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). Vol. 26. P. 279-285.
6. Muzychka S.A., Romanenko A.A., Piontkovskaja I.I. Conditional Random Field for morphological disambiguation in Russian // Компьютерная лингвистика и интеллектуальные технологии. 2014. С. 455-465.
7. Antonova A., Soloviev A. Using the method of conditional random fields for processing texts in Russian // Computer linguistics and intellectual technologies. 2013. P. 27–44.
8. Sorokin A., et al. MorphoRuEval-2017: an Evaluation Track for the Automatic Morphological Analysis Methods for Russian // Proceedings of the International Conference Dialogue 2017. 2017. V.1. Pp. 297-313.

9. Lyashevskaya O.N., et al. GRAMEVAL 2020 shared task: Russian full morphology and universal dependencies parsing // Proceedings of the International Conference Dialogue 2020. 2020. V.1. Pp. 553-569.
10. Chiche A., Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning aPproaches// Journal of Big Data. 2022. V.9. №1. Pp. 1-25.
11. Smirnov I.V. Intelligent text analysis based on methods of multi-level natural language processing. - M.: FRC IU RAS, 2023. - 356 p.
12. Anastasyev, D. G. Part-of-speech tagging with rich language description / D. G. Anastasyev, A. I. Andrianov, E. M. Indenbom // Computational linguistics and intellectual technologies: Based on the materials of the annual International Conference "Dialogue", Moscow, May 31 - 03 2017 / Russian State University for the Humanities. Vol. Issue 16 (23), Volume 1. - Moscow: Russian State University for the Humanities, 2017. - P. 2-13.
13. National Corpus of the Russian Language [Electronic resource]. – Access mode: URL: https://ruscorpora.ru/ (date of access: 20.11.2024).
14. Kopotev, M.V. National Corpus of the Russian Language / M.V. Kopotev, L.A. Yanda // Voprosy yazykoznaniya . – 2006. – No. 5. – P. 149–155.
15. Zhevnerovich, E.E. Corpus of texts in scientific research / E.E. Zhevnerovich // Proceedings of the II International Scientific and Practical Conference “Linguistics, Lingvodidactics, Linguistics and Cultural Studies: Current Issues and Development Prospects” (Minsk, March 1–2, 2018). – 2018. – P.25–32.
16. Nitsenko, A. V. Using the data of the National Corpus of the Russian language to remove the homonymy of the accusative and genitive cases within the paradigm of nouns / A. V. Nitsenko, V. Yu. Shelepov // Artificial intelligence: theoretical aspects, practical application: materials of the Donetsk international scientific round table. - Donetsk: FGBNU "IPII", 2023. - 252 p. - P. 137-141.
17. Lyashevskaya O. N., Sharov S. A. Frequency dictionary of the modern Russian language (based on the materials of the National Corpus of the Russian language). // M .: Azbukovnik, 2009. - 1112 p.

РЕЗЮМЕ

А. В. Ниценко, В. Ю. Шелепов

Об использовании семантической информации для снятия омонимии именительного и винительного падежа как элемента создания онтологии

Одной из ключевых задач в области обработки естественного языка является автоматическое снятие неоднозначности слов в текстах. Она заключается в выборе того значения многозначного слова, в котором оно употреблено в конкретном контексте. Неоднозначность, свойственная естественному языку, является серьёзным препятствием для компьютерного анализа текстов.

В статье описан метод автоматического разрешения омонимии между именительным и винительным падежами существительных, основанный на семантических связях слов, извлеченных из большого текстового корпуса. Эти семантические связи представлены в виде онтологии, включающей множество триплетов «субъект-предикат-объект».

В статье приведено описание предложенных методов и алгоритмов, рассмотрены примеры их работы. Разработанные правила для снятия неоднозначности были реализованы на языке программирования C++ в экспериментальной программе.

Учет семантической информации, заложенной в базе, позволил осуществлять снятие неоднозначности в случаях, когда без понимания смысла это сделать затруднительно.

RESUME

A.V. Nicenko, V. Ju. Shelepov

On the use of semantic information for disambiguation of the nominative and accusative cases as an element of creating an ontology

One of the key tasks in natural language processing is the automatic removal of ambiguity of words in texts. It consists of choosing the meaning of a polysemantic word in which it is used in a specific context. The ambiguity inherent in natural language is a serious obstacle to computer analysis of texts.

The article describes a method for automatic disambiguation between nominative and accusative cases of nouns based on semantic links between words extracted from a large text corpus. These semantic links are presented as an ontology that includes a set of "subject-predicate-object" triplets.

The article describes the proposed methods and algorithms, and considers examples of their operation. The developed rules for removing ambiguity were implemented in the C++ programming language in an experimental program.

Taking into account the semantic information embedded in the database made it possible to remove ambiguity in cases where it is difficult to do so without understanding the meaning.

Статья поступила в редакцию 21.06.2024.