СТАТИСТИЧЕСКОЕ РАЗРЕШЕНИЕ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ НЕОДНОЗНАЧНОСТИ В КОНТЕКСТАХ ДЛЯ ПРЕДМЕТНЫХ ИМЁН СУЩЕСТВИТЕЛЬНЫХ

STATISTICAL WORD SENSE DISAMBIGUATION IN CONTEXTS FOR NAMES OF PHYSICAL OBJECTS

Митрофанова О.А. *(alkonost-om@yandex.ru),* ***Паничева П.В.*** *(ppolin@yandex.ru),*
Санкт-Петербургский государственный университет;
Ляшевская О.Н. *(olesar@mail.ru), Институт русского языка им. В.В. Виноградова РАН*

В докладе обсуждаются результаты экспериментов по автоматизации процесса разрешения лексико-семантической неоднозначности слов. Эмпирическим материалом исследования являются примеры употребления предметных имён, извлечённые из Национального корпуса русского языка. Оцениваются оптимальные условия разрешения неоднозначности с учётом двух факторов: лексического наполнения контекстов и лексико-семантической разметки контекстов.

1. Постановка проблемы, цели и задачи исследования

Неоднозначность, свойственная естественному языку и проявляющаяся на различных его уровнях, является серьёзным препятствием для компьютерного анализа текстов. Разрешение лексико-семантической неоднозначности (наряду с морфологической и синтаксической) имеет особую важность в подготовке корпусов текстов, используемых системами автоматического понимания естественного языка. Выполнение этой процедуры представляет наибольшую сложность и зачастую требует ручной обработки текстов лингвистами-экспертами, в распоряжении которых находятся обширные словарные картотеки. Качество ручного разрешения неоднозначности оценивается как высокое, вместе с тем, желательно снизить трудоёмкость данной задачи за счёт использования специализированных компьютерных инструментов.

Итак, целью настоящего исследования является автоматизация процесса разрешения лексико-семантической неоднозначности текстов, что требует выполнения ряда задач, среди которых:

• подготовка компьютерного инструмента автоматического разрешения лексико-семантической неоднозначности слов в контекстах;

• обработка экспериментальных выборок, содержащих неоднозначные контексты;

• определение оптимальных условий, при которых качество разрешения лексико-семантической неоднозначности слов в контексте было бы высоким.

2. Исследовательские методы

Известны достаточно эффективные методы разрешения лексико-семантической неоднозначности в полуавтоматическом или автоматическом режиме [WSD 2006].[1] Методы первого типа предполагают использование компьютерных тезаурусов (WordNet, FrameNet) и формальных онтологий в качестве источников информации о значениях слов. Методы второго типа основываются на статистических данных о контекстном окружении слов, позволяющем разграничивать их употребление в различных значениях.

Применительно к материалу русского языка опробованы оба типа методов. Использование мощного электронного лексикографического ресурса (РуТез, семантический словарь НКРЯ) обеспечивает высокий уровень разрешения лексико-семантической неоднозначности [Лукашевич, Чуйко 2007; Кустова и др. 2006; Шеманаева и др. 2007]. Если же есть необходимость обойтись без словарной поддержки (например, в том случае, если обрабатываются тексты больших объёмов, а их лексический состав не покрывается имеющимися в распоряжении исследователей словарями), то предпочтение следует отдать статистическим методам. Достаточно надёжным является разрешение лексико-семантической неоднозначности на основе сравнения дистрибуций частеречных тегов контекстного окружения слов [Азарова, Марина 2006] и на основе лексических маркеров контекстов [Кобрицов и др. 2005]. Допустимо совмещение тезаурусного и статистического подходов к

[1] См. также материалы конференции SENSEVAL (www.senseval.org) и библиографию работ по WSD в материалах Corpora List (http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0512&L=corpora&D=1&F=&S=&P=2873).

разрешению лексико-семантической неоднозначности с опорой на словарную информацию о моделях сочетаемости слов [Кобрицов и др. 2007]. Можно предположить, что не менее (а возможно даже более) эффективной окажется статистическое разрешение неоднозначности с учётом дистрибуций лексико-семантических тегов в контекстах. Таких исследований на материале корпусов русского языка до нынешнего времени не проводилось. Эксперименты подобного рода впервые осуществлены в рамках обсуждаемого проекта.

В целях изучения возможностей статистического разрешения лексико-семантической неоднозначности в русскоязычных текстах предлагается адаптировать компьютерный инструмент автоматической классификации лексики таким образом, чтобы производилось сравнение неоднозначных контекстов с эталонными контекстами, представляющими реализацию того или иного значения слова. Классификация контекстов может быть основана как на сходстве их лексического состава, так и на сходстве лексико-семантических тегов для контекстных элементов (при наличии соответствующей разметки корпуса текстов).

3. Экспериментальный материал

Эксперименты по разрешению лексико-семантической неоднозначности проводились на материале Национального корпуса русского языка (НКРЯ)[2]. Были запланированы эксперименты двух типов, предполагавшие снятие неоднозначности а) на основе лексических маркеров значений слов в контекстах (тег леммы) и б) на основе лексико-семантической разметки контекстов (теги первого значения слова).

В качестве тестовых лексем выбраны предметные имена существительные. Известна филиация значений данных слов, фиксируемая в лексико-семантической аннотации НКРЯ. При описании значений анализируемых лексем использовалась структура значений слов в [ТСРЯ 1992]. Каждому значению соответствует особая комбинация тегов, принятых в системе разметки НКРЯ[3] (см. таблицу 1). Для рассматриваемых слов были сформированы выборки контекстов, присутствующих в НКРЯ (объёмы выборок см. в таблице 1). Очевидно, что анализируемые лексемы отличаются количеством значений, характером развития полисемии/омонимии, сте-

Значения	**Лексико-семантическая аннотация**	**Примеры**	**Число контекстов из НКРЯ**
дом			**3000**, из них
m1a. Жилое (или для учреждения) здание	r:concr t:constr top:contain	*Дом – новостройка.*	1694
m1b. Свое жильё	r:concr t:space	*Брать работу на дом.*	95
m2. Семья, люди, живущие вместе, их хозяйство	r:concr t:group pt:set sc:hum	*Мы знакомы домами.*	72
m3. Место, где живут люди, объединенные общими интересами, условиями существования	r:concr t:space der:shift der:metaph	*Общеевропейский дом.*	4
m4. Учреждение, заведение, обслуживающее какие-нибудь общественные нужды	r:concr t:org	*Дом культуры.*	292
m5. Династия, род	r:concr pt:set sc:hum	*Дом Романовых.*	1
диффузные значения *m1a/m1b, m1a/m2, m1b/m2* и пр.			842
орган			**834**, из них
m1. Клавишный духовой музыкальный инструмент, состоящий из труб, в к-рые нагнетается воздух	r:concr t:tool:mus	*Играть на органе.*	27
m2. Часть организма, имеющая определённое строение и специальное назначение	r:concr pt:partb pc:hum pc:animal hi:class	*Орган слуха.*	130
m2a. Орудие, средство	r:concr der:shift dt:partb	*Печать – активный орган пропаганды.*	9
m3. Государственное или общественное учреждение, организация	r:concr t:org hi:class	*Органы здравоохранения.*	660
m4. Печатное издание, принадлежащее какой-н. партии, организации, учреждению	r:concr t:media hi:class	*Академический орган.*	8
лук			**2200**, из них
m1. Огородное или дикорастущее растение сем. лилейных с острым вкусом луковицы и съедобными трубчатыми листьями	r:concr t:plant t:fruit t:food pt:aggr	*Репчатый лук.*	1600
m2. Ручное оружие для метания стрел в виде пружинящей дуги, стянутой тетивой	r:concr t:tool:weapon top:arc	*Стрельба из лука.*	600

Таблица 1. Филиация значений слов дом, орган, лук

[2] Публикации по НКРЯ: http://www.ruscorpora.ru/corpora-biblio.html

[3] Подробное описание системы тегов: http://www.ruscorpora.ru/corpora-sem.html

пенью связанности значений между собой. Необходимо отметить, что в рамках данного исследования используется трактовка неоднозначности, принятая в компьютерной лингвистике и допускающая условное приравнивание омонимичных коррелятов к многозначным словам [Рахилина и др. 2006]. Поэтому указанный материал для экспериментов по автоматическому разрешению неоднозначности является репрезентативным и позволяет получить результаты, соотносимые с разными условиями разрешения лексико-семантической неоднозначности.

Эксперименты по разрешению лексико-семантической неоднозначности проводились только для значений, представленных в НКРЯ достаточным количеством контекстов (например, из рассмотрения были исключены значения *m3* и *m5* для слова *дом*, значения *m2a* и *m4* для слова *орган*).

Было учтено, что в ряде контекстов регистрируется диффузность значений исследуемых лексем: например, *дом – m1a* (строение) vs. *дом – m1b* (личное пространство, которое часто физически оказывается вовсе не домом, а комнатой или квартирой, ср. отыменные наречия *дома*, *домой*). Подобные контексты были проанализированы отдельно.

4. Постановка экспериментов

Разрешение лексико-семантической неоднозначности слов в корпусе рассматривается как задача распознавания образов. В качестве экспериментальной выборки используется набор контекстов, в которые вручную введены лексико-семантические теги, соответствующие значениям исследуемых лексем. Из экспериментальной выборки контекстов для той или иной лексемы автоматически формируются образы – эталонные классы контекстов, иллюстрирующие употребление слова в каком-либо одном значении. В образ попадают контексты, отобранные случайно. Оставшиеся тестовые контексты (все или часть из них) автоматически сравниваются с образами и распределяются по группам в соответствии со значениями, в этом случае априорная лексико-семантическая информация об исследуемых лексемах не используется: значение лексемы определяется автоматически. Тем самым, разрешение неоднозначности предполагает автоматическую классификацию контекстов употребления лексемы в разных значениях. Данная процедура требует представления экспериментальной выборки как векторного пространства, где каждый контекст преобразуется в вектор. Близость контекста употребления слова в каком-либо значении к тому или иному образу оценивается с помощью трёх мер расстояния: меры Евклида (*Eucl*), меры Хемминга (*Hm*) и значения косинуса угла между контекстными векторами (*Cos*). Данные меры имеют некоторые особенности. Если мера Хемминга линейна (и она аккумулирует разницы по координатам для двух точек), то мера Евклида отражает квадратичную зависимость расстояния между точками от разниц по их координатам (она аккумулирует квадраты разниц по координатам). В обоих случаях на результат влияют как раз большие разницы, это влияние слабее для меры Хемминга и сильнее для меры Евклида. В отличие от меры Евклида и меры Хемминга, мера косинуса менее чувствительна к большим разницам по отдельным координатам и не зависит от длин векторов.

Для исследуемых слов была проведена серия экспериментов с различными по объёму эталонными классами и тестовыми выборками контекстов, с изменением меры близости, с опорой на лексические маркеры значения в контексте либо на лексико-семантические теги. Во всех экспериментах объём контекста не ограничивался каким-либо окном. Результаты автоматической классификации контекстов сравнивались с результатами ручной разметки значений слов в контекстах.

5. Компьютерное обеспечение экспериментов

В экспериментах использовался компьютерный инструмент автоматической классификации лексики [Митрофанова и др. 2007], адаптированный для разрешения неоднозначности слов в контексте. Реализован алгоритм классификации с учителем. Программное обеспечение разработано П.В. Паничевой на языке Python. В ходе работы программы производятся следующие процедуры.

Во-первых, производится подготовительная обработка экспериментальных выборок контекстов. В выборке определяются те контексты, в которых значение лексемы может быть идентифицировано однозначно. Вычисляется количество имеющихся контекстов для каждого из значений исследуемой лексемы. Для значений с достаточным числом контекстов случайным образом формируется тестовая выборка и не пересекающаяся с ней обучающая выборка (эталонный класс). Для дальнейшей работы программы необходимо, чтобы для каждого значения были сформированы два файла, в которых приведены тестовая выборка и эталонный класс.

Во-вторых, осуществляется процесс машинного обучения. Для исследуемых значений программа производит обработку файла с эталонными классами контекстов, в ходе которой формируется образ значения. Из эталонных контекстов извлекается лексическая информация, тем самым, в образ значения включаются все лексемы, встретившиеся в эталонных контекстах, с учётом частоты их встречаемости. На выходе процедуры

формируется статистические образы значений анализируемого слова, представленные словарём, в котором указаны лексемы и их относительная частота. Таким образом, если обучающая выборка для одного из значений слова *лук* составляла бы 100 контекстов, и в них 50 раз встретилась лексема *резать* и 30 раз встретилась лексема *морковь*, то в статистическом образе этого значения глагол *резать* имел бы показатель частотности 0,5, а существительное *морковь* – 0,3. Итак, образ значения можно рассматривать как вектор в векторном пространстве, координаты которого определяются частотными показателями соответствующих лексем, встретившихся в обучающей выборке контекстов для этого значения. В экспериментах с учётом лексико-семантической информации статистический образ формируется аналогичным путём, однако координатами в векторном пространстве служат не слова, а лексико-семантические теги слов, выступающих в качестве контекстного окружения исследуемых лексем.

Далее программа, прошедшая обучение, обрабатывает тестовые выборки контекстов. Для этого каждый контекст также рассматривается как вектор в векторном пространстве, и вычисляется мера расстояния данного контекста по отношению к векторам, представляющим образы значений. Выбирается образ значения, который оказывается наиболее близким к образу анализируемого контекста, в итоге, этому контексту присваивается соответствующее значение. При проверке результатов классификации для каждого из значений вычисляется количество правильных решений – тех случаев, когда автоматическая оценка значения, реализованного в контексте, совпадает со значением, назначенным вручную и отражённым в лексико-семантических тегах исследуемой лексемы.

6. Результаты экспериментов по автоматическому разрешению лексико-семантической неоднозначности слов в контекстах

6.1. Иллюстрация результатов компьютерной обработки контекстов

В ходе экспериментов обрабатываемым неоднозначным контекстам для предметных имён существительных автоматически приписывалось то или иное значение. Так, в таблице 2 приведены некоторые примеры анализа контекстов слова *дом*.

Контексты (в квадратных скобках указан номер контекста в корпусе)	**Исходное значение**	**Распознанное значение**	***Cos***
[649] Я помню всю эту чепуху детства, потери, находки, то, как я страдал из-за него, когда он не хотел меня ждать и шёл в школу с другим, и то, как передвигали *дом* с аптекой, и ещё то, что во дворах всегда был сырой воздух, пахло рекой, и запах реки был в комнатах, особенно в большой отцовской, и, когда шёл трамвай по мосту, металлическое бренчание и лязг колёс были слышны далеко.	*m1a*	*m1a*	0,650
[3004] Уже два года, как Таня ушла *из дому* и жила по разным местам, у новых приятелей, – то в мастерской знакомого художника на Шаболовке, то на пустующей зимней даче чьих-то родственников под Звенигородом, то в служебной квартире подружки, работавшей техником-смотрителем на Молчановке...	*m1b*	*m1b*	0,438
[957] Все подъезды в этом *доме* – со двора.	*m1a*	*m4*	0,288
[2130] Домишко рядом с *домом* подполковника.	*m1a*	*m2*	0,099
[3042] Пришлось Анну вернуть в *дом*, вскоре и Катю поселили.	*m1b*	*m4*	0,410

Таблица 2. Примеры компьютерной обработки контекстов употребления слова дом

Примеры [649] и [3004] проанализированы верно, тогда как примеры [957], [2130] и [3042] интерпретируются неточно. Вероятно, ошибочные решения связаны с недостаточностью контекстного окружения для идентификации значений.

Результаты автоматического разрешения неоднозначности дополняются информацией о контекстных маркерах лексических значений исследуемых слов в контекстах (см., например, таблицу 3).

Значения	**Лексические маркеры**
m2. Часть организма…	*порок, врождённый…*
m3. …Учреждение, организация…	*учреждение, самоуправление, начальник, местный, правоохранительный…*

Таблица 3. Примеры лексических маркеров значений слова орган *в контекстах*

6.2. Оптимальные условия автоматического разрешения лексико-семантической неоднозначности слов в контекстах

Данные, полученные в процессе исследования, свидетельствуют о следующих фактах.

Во-первых, наилучшие результаты разрешения лексико-семантической неоднозначности на основе лексических маркеров (в среднем 85% правильных решений, в отдельных случаях до 95% правильных решений) могут быть получены при использовании в качестве меры расстояния значения косинуса угла между контекстными векторами (см. таблицу 4) .

Мера	***Eucl***	***Hm***	***Cos***
Точность *(p)*	0,45	0,65	0,85

Таблица 4. Точность результатов автоматического разрешения лексико-семантической неоднозначности слов в контекстах с использованием различных мер

Во-вторых, успешность разрешения лексико-семантической неоднозначности находится в прямой зависимости от частотности контекстов с тем или иным значением слова в экспериментальной выборке. Частотность значения сказывается на чёткости формируемого эталонного класса. Эталонные классы для частотных значений являются более чёткими, чем классы для значений с умеренной частотой. Так, для слова *орган* высокочастотное значение *m3* распознаётся лучше, чем низкочастотное значение *m1* и значение *m2* с умеренной частотой. По всей видимости, хороших результатов распознавания можно достигнуть при наличии не менее 100 контекстов употребления слова в экспериментальной выборке.

В-третьих, изменение объёма эталонного класса (S = 15, 55, 75, 100, 200, 500, … полная выборка за исключением тестовых контекстов) также оказывает существенное влияние на качество разрешения лексико-семантической неоднозначности. При предельных объёмах эталонных классов качество распознавания оказывается низким, поскольку в эталонном классе малого объёма недостаточно контекстов для фиксации признаков употребления слова в том или ином значении, а в максимально широком эталонном классе велика доля случайных признаков, не сопряжённых с конкретным значением (см., например, таблицы 5 и 6).

Объём эталонного класса (S)	**Точность (p)**	**Объём эталонного класса (S)**	**Точность (p)**	**Объём эталонного класса (S)**	**Точность (p)**
15	0,63	75	0,77	200	0,56
55	0,80	100	0,8	полная выборка	0,77

Таблица 5. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова орган *в контекстах с использованием меры* Cos *и с учётом объёма эталонного класса*

Объём эталонного класса (S)	**Точность (p)**	**Объём эталонного класса (S)**	**Точность (p)**	**Объём эталонного класса (S)**	**Точность (p)**
100	0,78	500	0,83	полная выборка	0,73

Таблица 6. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова лук *в контекстах с использованием меры* Cos *и с учётом объёма эталонного класса*

6.3. Сравнение результатов автоматического разрешения лексико-семантической неоднозначности на основе лексических маркеров и лексико-семантических тегов

Была проведена серия экспериментов для сравнения эффективности автоматического разрешения лексико-семантической неоднозначности слов на основе лексических маркеров, выявляемых в их контекстах, и лексико-семантических тегов их контекстного окружения. Например, в таблице 7 приведены некоторые контексты, иллюстрирующие употребление слова *лук* в значениях *m1* и *m2*, а также результаты их компьютерной обработки в двух режимах (объём тестовых выборок – 20 контекстов, объём эталонных классов – 500 контекстов, мера *Cos*).

	Распознавание на основе лексических маркеров		**Распознавание на основе лексико-семантических тегов**	
Контексты (в квадратных скобках указан номер контекста в корпусе)	**Распознанное значение**	**Cos**	**Распознанное значение**	**Cos**
исходное значение m1				
[2379] Помню хлеб с изюмом, с *луком*, с какими-то кореньями.	*m1*	0,572	*m1*	0,786
[1578] Щавель –300 г, огурцы – 50 г, *лук* зелёный – 30 г, яйца – 1 шт., сметана – 30 г, сахар – 10 г, укроп.	*m1*	0,653	*m1*	0,569
[193] Начинают принимать *лук*, капусту – гляди в оба глаза.	*m2*	0,502	*m1*	0,514
исходное значение m2				
[235] Одни тугие *луки*, над которыми несколько человек справиться не могли, «играючи» натягивали, другие толстенные железные полосы вокруг шеи врага скручивали, третьи возы через броды на себе перетаскивали, ядра через самые широкие реки запросто перебрасывали.	*m2*	0,533	*m2*	0,550
[1120] Знаешь, есть восточное присловье, что, если человек стреляет из *лука*, он никогда не попадет в мишень, если стрела не пробьет одновременно его сердце.	*m2*	0,543	*m2*	0,538
[1863] Не имев совершенного успеха в намерении взбунтовать тушинский стан и боясь мести гетмана, Марина, в одежде воина, с *луком* и тулом за плечами, [11 февраля] ночью, в трескучий мороз ускакала верхом к мужу, провождаемая только слугою и служанкою.	*m1*	0,507	*m2*	0,609

Таблица 7. Примеры компьютерной обработки контекстов употребления слова лук

Оценки точности автоматического разрешения лексико-семантической неоднозначности при заданных условиях приведены в таблице 8.

	Точность (p)		**Среднее ($p_{ср}$)**
	лук *(m1)*	**лук *(m2)***	
Распознавание на основе лексических маркеров	0,75	0,9	0,83
Распознавание на основе лексико-семантических тегов	0,75	0,95	0,85

Таблица 8. Точность результатов автоматического разрешения лексико-семантической неоднозначности слова лук *в контекстах на основе лексических маркеров и лексико-семантических тегов*

В подавляющем большинстве случаев распознавание на основе лексических маркеров и на основе лексико-семантических тегов приводит к одинаково правильным решениям (см. примеры [2379], [1578], [235], [1120] в таблице 7). Вместе с тем, результаты разрешения лексико-семантической неоднозначности по тегам часто оказываются лучше, чем результаты, полученные при использовании лексических маркеров (ср. значения меры косинуса для примеров [2379] и [235]). Были зарегистрированы контексты, показывающие незначительное снижение значения меры косинуса (ср. примеры [1578] и [1120]), однако это не влияет на качество распознавания при переходе от лексических маркеров к тегам. Важно, что в ходе анализа экспериментальных данных удалось получить подтверждение гипотезы о том, что при разрешении неоднозначности на основе лексико-семантических тегов удаётся улучшить результаты идентификации значений слов в контексте и избежать ошибочных решений (см. примеры [193] и [1863]). Среди причин, вызывающих неудачи при разрешении лексико-семантической неоднозначности, можно указать недостаточность (вплоть до полного отсутствия) диагностических маркеров значения в чрезмерно коротких контекстах (см. пример [193]) или, наоборот, в слишком широких контекстах (см. контекст [1863]). Как правило, значение меры косинуса в этих случаях удерживается около показателя 0,5. Возможный путь корректировки результатов автоматического анализа связан с дополнительным использованием других мер расстояния.

6.4. Анализ контекстов с диффузными значениями

Наряду с экспериментами по автоматической обработке потенциально однозначных контекстов употребления слов было произведено разрешение лексико-семантической неоднозначности в контекстах с диффузными значениями, а также сравнение результатов ручного и компьютерного анализа. В таблице 9 приведены примеры некоторых диффузных контекстов слова *дом*, указывающие на возможность выбора доминирующего значения в паре по итогам компьютерного анализа.

Контексты (в квадратных скобках указан номер контекста в корпусе)	Диффузные значения	Распознанное значение	**Cos**
[337] А в *доме* у Ёжика топилась печь, потрескивал в печи огонь, а сам Ёжик сидел на полу у печки, помаргивая, глядел на пламя и радовался.	*m1a*/*m1b*	*m1a*	0,429
[2983] Семён на портфель и не взглянул, а заточку аккуратно обтёр кухонной тряпкой, предусмотрительно им захваченной *из дому*, засунул инструмент в рукав, под часовой ремень, и вышел из двора той новой походкой, негнущейся и манекенной, которая образовалась у него после больничного излечения...	*m1a*/*m1b*	*m1b*	0,541
[3214] Родственники у Ливии все как один люди практичные, богатые и важные, хоть и не без вывертов; кажется, единственный человек, который уважает её в этом *доме*, – это ее дворецкий, Трефль.	*m1b*/*m2*	*m2*	0,452

Таблица 9. Примеры компьютерной обработки сложных случаев употребления слова дом *в контекстах*

В дальнейшем условия эксперимента были изменены, дополнительно сформированы эталонные классы для диффузных значений типа *m1a*/*m1b*, *m1a*/*m2*, *m1b*/*m2* и пр.

7. Выводы и перспективы развития исследования

В результате исследования была проведена модернизация компьютерного инструмента автоматической классификации лексики и введение специализированного режима его работы, позволяющего автоматически классифицировать неоднозначные контексты употребления слов в соответствии с присущими им значениями. Был реализован алгоритм классификации объектов с учителем и процедуры автоматической обработки контекстов с опорой на лексическое наполнение контекстов, а также с учётом лексико-семантических тегов, приписываемых контекстному окружению слов.

Были проведены серии экспериментов по автоматическому разрешению неоднозначности контекстов употребления предметных имён существительных с различной семантической структурой. Данные слова характеризуются разным числом значений, отличающихся по частотности и по степени самостоятельности. Это позволило получить обширные экспериментальные данные на русскоязычном материале и оценить оптимальные условия, обеспечивающие достаточно высокое качество разрешения семантической неоднозначности слов в контекстах (от 85% и выше).

Оптимальными можно признать следующие условия разрешения лексико-семантической неоднозначности слов в контекстах:

• высокий объём экспериментальной выборки;

• наличие в выборке не менее 100 контекстов употребления слова в исследуемом значении;

• объём эталонного класса около 500 контекстов;

• оценка близости контекстов к эталонному классу с использованием значения косинуса угла между контекстными векторами;

• возможность снятия неоднозначности на основе лексических маркеров значения слова в контексте либо на основе лексико-семантических тегов его контекстного окружения.

В ходе экспериментов нашла подтверждение гипотеза о большей эффективности разрешения лексико-семантической неоднозначности с опорой на лексико-семантическую разметку корпуса текстов.

Продолжение исследования предполагает проведение экспериментов по разрешению семантической неоднозначности:

• на обширном корпусном материале (увеличение экспериментальной группы лексем, использование большеобъёмных экспериментальных выборок контекстов из корпуса);

• с оценкой контекста на основе комбинированных признаков (например, с учётом как лексических, так и лексико-семантических данных, с вычислением оптимальных весовых коэффициентов в контекстах и пр.);

• с изменением ширины контекстного окна (в предыдущих экспериментах рассматривались контексты в полном объёме, предлагается сужать границы контекстов и варьировать протяжённость обрабатываемых фрагментов контекстов);

• с детальным анализом диффузных контекстов употребления лексем в сопряжённых значениях (определение доминирующего значения: например, *стакан с водой* (*стакан* – «вместилище») vs. *стакан воды* (*стакан* – «мера+вместилище»);

• с проверкой ряда статистических гипотез об условиях разрешения лексико-семантической неоднозначности лексем в корпусах текстов.

Список литературы

1. Азарова И.В., Марина А.С. Автоматизированная классификация контекстов при подготовке данных для компьютерного тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 13–17.

2. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка // Интернет–математика 2005: Автоматическая обработка веб-данных. М.: 2005. С. 38–57.

3. Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю. Снятие семантической многозначности глаголов с использованием моделей управления, извлечённых из электронных толковых словарей // URL: http://download.yandex.ru/IMAT2007/kobricov.pdf

4. Кустова Г.И., Рахилина Е.В., Ляшевская О.Н., Шеманаева О.Ю. Семантическая разметка и семантические фильтры для Национального корпуса русского языка // Труды международной конференции «Корпусная лингвистика–2006». СПб.: 2006. С. 209–218.

5. Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний // Интернет–математика 2007: Сборник работ участников конкурса. Екатеринбург: 2007. С. 108–117.

6. Митрофанова О.А., Мухин А.С., Паничева П.В. Автоматическая классификация лексики в русскоязычных текстах на основе латентного семантического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2007». М.: 2007. С. 413–421.

7. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2006». М.: 2006. С. 445–450.

8. ТСРЯ – Ожегов С.И., Шведова Н.Ю. Толковый словарь русского языка. М., 1992.

9. Шеманаева О.Ю., Кустова Г.И., Ляшевская О.Н., Рахилина Е.В. Семантические фильтры для разрешения многозначности в национальном корпусе // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007. С. 582–587.

10. WSD – Word Sense Disambiguation: Algorithms and Applications / Eds. E. Agirre, Ph. Edmonds. Springer: 2006.