Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка

Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.

ВИНИТИ РАН
neuralman@yandex.ru, olesar@mail.ru,
shemanaeva@yandex.ru

Аннотация

Задачу снятия лексико-семантической омонимии (word-sense disambiguation) [Hirst 1986; Ide, Veronis 2002; Stevenson 2003 и др.] в семантически размеченных текстах предлагается решать с помощью поверхностных фильтров, или жестких правил-шаблонов (Weiss 1973). Эти правила дают наибольшую точность результатов, но в то же время считаются наиболее трудоемкими. Наш метод комбинирует автоматическое построение БД высокочастотных устойчивых коллокаций с их последующим (полу)ручным аннотированием. В качестве исходного материала выбран корпус публицистики, поскольку в таких текстах особенно велика доля языковых штампов (*вступить в силу, уровень жизни* и т. д.). В настоящее время построена система фильтров, основанная на 4500 частотных устойчивых сочетаниях слов. Эксперименты по применению поверхностных фильтров для разных подкорпусов показывают уменьшение омонимии от 3,3 до 6%.

1. Введение и обзор ключевых новейших работ по исследуемой тематике

Большинство систем разрешения семантической неоднозначности, разрабатывавшихся до 1980 гг., были основаны на использовании

правил, созданных вручную. Такие правила представляли собой операцию условного выбора, когда для каждого допустимого типа контекста словоупотребления слова W выбиралось одно значение из словаря. Выбор правильного значения осуществлялся на основе анализа семантических ограничений, налагаемых на контекст рассматриваемого слова (см. Weiss 1973; Kelly, Stone 1975; Small, Rieger 1982; подробный обзор см. в Кобрицов 2004а). С. Вайс предложил использовать правила двух типов: *общие контекстные правила* и *правила-шаблоны*. Общее контекстное правило задает выбор определенного значения слова, если это слово употребляется рядом с некоторым конкретным словом. Например, если со словом *type* 'тип, печатать на машинке' в предложении появляется слово *print* 'печатать', то значением *type* скорее всего будет 'печатать'. Правила-шаблоны жестко определяют ближайший контекст слова. Например, если слово *of* 'из, показатель притяжательности' появляется непосредственно справа от слова *type*, то значением последнего будет скорее всего 'множество, тип чего-либо'. После серии проведенных тестов своей системы Вайс обнаружил, что использование правил-шаблонов дает гораздо более точные результаты. Точность разрешения неоднозначности построенного алгоритма составила порядка 90 %. Проанализировав случаи неправильного выбора, Вайс выяснил, что в основном это случаи идиоматического использования слова.

Вайс создал правила всего для 5 слов. Как стало очевидно после работ Келли и Стоуна и Смолла и Ригера, при использовании таких систем для больших словарей трудозатраты на создание правил становятся слишком большими: «описание … слова "throw" занимает уже 6 полных страниц … это много, однако должно быть в 10 раз больше» (Kelly, Stone 1975).

В середине 1980 гг. акцент в исследованиях по разрешению семантической неоднозначности сместился от вручную созданных индивидуальных правил для каждого слова к автоматически порожденным правилам на основе данных, извлекаемых из корпусов текстов. Можно выделить четыре основных вида данных, которые определили новые методы дизамбигуации:

1. данные словарей (Lesk 1988; Wilks et al. 1990; Gutrie et al. 1991; Demetriou 1993; Demetriou, Atwell 2001; Stevenson, Wilks 2002),

2. вручную размеченный тренировочный корпус (Black 1998; Hearst 1991; Yarowsky 1995),
3. переводные словари и переводные корпуса (Dagan et al. 1991; Gale et al. 1992),
4. тезаурусы (Masterman 1957; Patrick 1985; Yarowsky 1992, Sussna 1993).

В первом случае для разрешения семантической неоднозначности в качестве корпуса используются толкования словаря, как правило, Longman Dictionary of Contemporary English (LDOCE). В системе Й.Уилкса (Wilks et all. 1990) для каждого слова W берется его словарное толкование, затем из него вычеркиваются все семантически нерелевантные слова (предлоги, союзы и т.п.) – в результате получается множество метаслов $\{w_1,..,w_n\}$. После этого такое множество метаслов сравнивается с толкованиями других слов: если слово w_i из толкования рассматриваемого слова W присутствует в нескольких других толкованиях вместе с метасловом v_j, то (основываясь на допущении о семантическом родстве вместе встречающихся слов) алгоритм устанавливает дополнительную связь между словом W и v_j. Таким образом, помимо семантической связи между W и w_i, которая устанавливается из-за того, что w_i присутствует в толковании рассматриваемого слова W, устанавливается еще и связь между W и v_j, именно так и происходит "расширение" толкования. По окончании работы для рассматриваемого слова W^N образуется семантическая сеть, связи в которой отражают некоторую степень семантического родства. Значение слова в тексте устанавливается по вхождению элементов контекста слова W в ту или иную семантическую сеть $W^1 \ldots W^N$. Точность работы алгоритма Уилкса составила 53%, а при определении правильного значения на более общем уровне дробности значений точность работы оказалась равной 85%.

Другой метод автоматического разрешения неоднозначности основан на предварительном выборе правильного значения слов в некотором корпусе вручную – полученные данные впоследствии используются для "обучения" алгоритма. Первое исследование такого рода на большом материале было предпринято Блэк (Black 1988). Работая с дробными значениями LDOCE, она взяла пять многозначных слов, у которых различалось по крайней мере три значения внутри одной морфологической категории, и для каждого слова вручную разметила 2000 употреблений слова. Затем Блэк случайным образом разделила каждое множество употреблений на

"тренировочную группу", состоящую из 1500 употреблений и тестовую группу, в которую вошли оставшиеся 500. Блэк придумала три процедуры снятия неоднозначности, которые были последовательно примены к пяти тренировочным корпусам. Каждая процедура использует разные свойства тренировочных групп: одна работает с семантическими категориями из LDOCE, две остальных автоматически порождают правила выбора значения на основе текста толкования (метод, похожий на правила Вайса). Затем была измерена точность каждой процедуры для каждого тестового множества. Метод снятия семантической многозначности с использованием семантической классификации показал точность в 45%, два других казались примерно одинаковыми, показав 72% и 75%. Система М. Харст (Hearst 1991) сочетает в себе "управляемое" (supervised) обучение на размеченном вручную подкорпусе употреблений слова и "самостоятельное обучение" на неразмеченных употреблениях. Здесь алгоритм проводит выбор правильного значения слова и собирает из контекста такие же типы лексических и грамматических ключей, которые извлекались во время управляемого обучения. Неуправляемое обучение показало приемлемые результаты, несмотря на то, что для начала работы системы необходимо было иметь довольно большое количество вручную размеченных слов. Харст сообщила, что точность работы системы для 6 слов варьировалась от 73% до 100%, хотя идеальная работа была отмечена только для одного слова. Наконец, в недавней работе Яровски (Yarowsky 1997) сообщалось, что построенная им система, работающая на подобных методах, достигает точности 97%. Как видно, построение таких систем опять же требует достаточно большой работы по ручной разметке множества употреблений для каждого многозначного слова.

Идея третьего метода, основанного на двуязычных словарях и параллельных корпусах, такова: пусть предложение на исходном языке содержит слова, которые могут переведены разными способами. Система генерирует все комбинации таких переводных слов (заметим, что здесь возможен комбинаторный взрыв). Затем каждая такая комбинация рассматривается в корпусе текстов на языке перевода и проводится подсчет частотности появления такой комбинации в корпусе. Наиболее частотная комбинация определяет правильный перевод слова и правильное значение, отражающее соответствующий перевод. В работе Dagan et al. 1991 этот алгоритм тестировался на двух переводах: немецкий в английский и иврит в английский. В общем была проведена попытка разрешения 159

употреблений слов (105 на иврите, 54 немецких), из которых 54 были отброшены, так как в корпусе на конечном языке было недостаточно примеров употреблений этих слов. Для оставшихся 105 (73 ивритских и 32 немецких) точность разрешения составила 75% для разрешения в паре немецкий-английский и 92% для пары иврит-английский. Точность системы для пары немецкий-английский оказалась ниже вследствие небольшого числа употреблений возможных переводных слов в английском корпусе.

Тезаурус Роже, а также семантические сети типа WordNet, являются очень удобными источниками для построения систем снятия семантической неоднозначности. Так, М. Суссна (Sussna 1993) использовал сеть WordNet, чтобы вычислять семантическое расстояние между двумя словами в сети. Для этого он присвоил определенный вес каждому отношению между синсетами в семантической сети. Величина веса, приписанного отношению, отражает семантическую близость между синсетами, связанными данным отношением. Например, отношение синонимии получило наивысший вес, тогда как отношение антонимии получило наименьший вес. Семантическое расстояние между двумя синсетами можно вычислить, суммировав все веса, присвоенные отношениям, через которые лежит кратчайший путь между двумя этими синсетами. Суссна испытывал свой алгоритм в разных конфигурациях. Основные параметры, которые он менял, были размер контекста при выборе значения слова и количество слов, для которых одновременно осуществлялось разрешение семантической неоднозначности. При выборе значения для более одного слова одновременно система Суссны проверяла каждую комбинацию значений. Это часто приводило к экспоненциальному росту количества возможных комбинаций. Тестирование системы было проведено на корпусе газетных материалов TIME. Из этих документов было отобрано 319 употреблений рассматриваемых слов и для них Суссна вручную снял семантическую многозначность. Алгоритм выбрал правильное значение с точностью в 56%. Сообщалось, что контекст из 41 слова дал наиболее высокую точность при разрешении неоднозначности. Помимо этого, при одновременном выборе значений повысилась точность, однако из-за роста числа комбинаций значений в системе было введено ограничение на обработку не более десяти многозначных слов одновременно.

Один из самых точный на сегодняшний день алгоритм разрешения семантической неоднозначности был построен Д. Яровски на основании данных тезаурусов Роже и Grolier Multimedia Encyclopedia. Система использует 1042 семантических категории, на которые разделены все слова в тезаурусе Роже. Это достаточно общие категории, описывающие такие области, как *машины* или *животные-насекомые*. Для определения, к какой семантической категории относится данное многозначное слово, используется множество слов-"ключей" для конкретной категории (для каждой категории свое множество), полученных из грамматически размеченной Энциклопедии Grolier. Чтобы получить такое множество слов-"ключей", необходимо для каждого слова из этой категории проанализировать все употребления этого слова в толкованиях словаря Grolier, и для каждого употребления выбрать его контекст (100 слов вокруг данного употребления). Например, в категорию "инструмент-машина" входят 348 слова, которые встречаются в Grolier 30924 раза. Для каждого из этих употреблений собраны их контексты. Для каждого слова из набора контекстов определялась частота его появления во всех контекстах. Затем полученная частота сравнивалась с частотой появления этого слова в тексте всей энциклопедии, и слову присваивался определенный вес, основанный на сравнении этих частот. Словами-"ключами" для данной семантической категории становились слова с наивысшим весом. Яровски сообщил, что для каждой категории было получено около 3000 слов-"ключей". Для тестирования Яровски провел предварительное обучение своей системы на 12 многозначных словах: вручную было размечено несколько сотен употреблений каждого из этих слов. Точность работы алгоритма варьировалась, но в среднем составила 92%.

Итак, несмотря на то, что точность работы многих современных систем достигает 90% и выше, в основном они могут проводить выбор правильного значения лишь для нескольких слов. Если же говорить об уменьшении лексико-семантической омонимии для всего текста в целом, то она снижается лишь на сотые, если не тысячные доли процента. Ряд алгоритмов отслеживает многозначность только на уровне крупных групп значений, и не позволяет провести различие между отдельными словарными значениями многозначного слова. Кроме того, некоторые системы преследует проблема комбинаторного взрыва при переборе гипотез. Имея в своем распоряжении значительно большие электронные ресурсы, проверенные временем методы, исследователи 1990-2000

гг. несомненно улучшили результаты предшественников, но сейчас наблюдается некоторая стагнация в смысле новых идей. Работы самого последнего времени (Demetriou, Atwell 2001; Stevenson, Wilks 2002; Pedersen 2005) стремятся преодолеть известные проблемы и применяют в основном сочетание разных методов.

2. Идея исследования

2.1. От словоцентричных правил к устойчивым коллокациям

Описанные выше подходы являются по преимуществу словоцентричными – исследователи стремятся построить правила (автоматически или вручную, неважно), которые *исчерпывающим образом* определяют выбор значения для некоторого *слова*. Однако сейчас акцент в машинно-ориентированном анализе семантики текстов перемещается с отдельного слова на устойчивые коллокации. Например, автоматическое тезаурирование в системе Sketch Engine (Kilgarriff et al. 2004; http://www.sketchengine.co.uk) построено на синтаксически аннотированной автоматическим способом БД коллокаций (80 млн. записей). БД устойчивых коллокаций используются также для создания словарей, выравнивания параллельных корпусов (LEXICOM 2005), обучения языку как иностранному, автоматическому реферированию, анализа малапропизмов, лингвистической стеганографии (Bolshakov 2005) и др.

Идея настоящего проекта родилась в ходе работ по "ручному" снятию морфологической омонимии в Корпусе современного русского языка (www.ruscorpora.ru). Когда разметчики, накопив опыт работы по снятию омонимии в художественной литературе, перешли к обработке газетных и других нехудожественных текстов, выяснилось, что в них из статьи в статью или, например, в мемуарах одного автора довольно часто повторяются одни и те же обороты:

На самом деле, все отставки были политически мотивированы.

Взять, например, **идею с** референдумом **по поводу** отставки министра обороны.

А вот с культурой в нашем Отечестве **не все в порядке**.

[примеры из Национального корпуса]. Заметим, что Д.Н. Шмелев еще в 1964 году отмечал растущую роль многословных оборотов в современном русском языке: "Некоторые из слов… все активнее употребляются для выражения разного рода отношений между предметами и явлениями, обозначенными другими словами. В связи с этим их функция в предложении становится близкой функции предлогов, союзов. Ср. такое употребление слов *сфера, лицо, дух, мера* и т. п.: *перед лицом военной угрозы, в сфере распределения материальных благ, со стороны общества, в расчете на массового потребителя, в духе взаимопонимания, по мере того как, в связи с этим, что* и т.п." [Шмелев 2002:87-88]. Очевидно, что учет подобных конструкций - устойчивых коллокаций может быть полезен не только при снятии морфологической омонимии, но также в "малом" синтаксисе (shallow parsing)[1] и при разрешении лексической многозначности в семантической разметке. Кроме того, информация о вхождении слова в некоторый устойчивый оборот может представлять самостоятельную ценность для пользователя корпуса.

2.2. Задачи исследования

В ходе исследования были поставлены следующие задачи:

1. автоматически выделить списки частотных коллокаций на базе большого корпуса публицистических текстов; на их основе создать базу данных коллокаций, содержащую морфологическую и семантическую разметку,
2. создать правила для выявления списка лингвистически релевантных устойчивых коллокаций с применением статистических и лингвистических методов,
3. снять лексико-семантическую омонимию в выделенных коллокациях (вручную, но с применением полуавтоматических процедур оптимизации работы с морфологическими и семантическими классами),
4. оценить эффективность работы фильтров.

[1] В частности, это может уменьшить вероятность комбинаторного взрыва при переборе синтаксических гипотез.

2.3. Новизна исследования

Здесь можно выделить несколько аспектов. Хотя устойчивые коллокации давно используются при анализе русскоязычных текстов (см., например, исследования Большаков, Галисия-Аро 2003; Борисова 1995; Добровольский 2003; Копотев 2004 и др.), они впервые применяются к снятию семантической омонимии. Более того, нам неизвестны аналогичные работы по снятию семантической омонимии в англоязычных текстах, поскольку считается, что этот подход требует больших человеческих ресурсов. Однако, если оценивать нашу работу по соотношению объемов ручной обработки и эффекта от работы системы (уменьшению процента многозначности в текстах), то это соотношение оказывается гораздо ниже, чем для работ, построенных на отдельных словоцентричных правилах. Наконец, мы получаем базу лингвистически релевантных частотных устойчивых коллокаций (п. 2), к достоинствам которой можно отнести: большой размер, метод получения (с участием эксперта-человека), возможность применения этих данных к другим методам семантической разметки и к другим видам парсинга (морфологическому и синтаксическому).

3. Описание методов, алгоритмов и экспериментов

3.1. Исходные данные. Национальный корпус русского языка. Лексико-семантическая информация в корпусе

Семантическая разметка Корпуса современного русского языка содержит информацию о принадлежности лексемы к одному или нескольким традиционным лексико-семантическим классам, таким как "глаголы движения", "каузативные глаголы", "части тела", "имена деятеля" и т. п. В настоящее время семантический разбор получают имена существительные, прилагательные, числительные, местоимения, глаголы и наречия (подробнее о разметке см. [Кустова et al. 2004; Кустова et al. (в печати)], а также документацию на сайте http://www.ruscorpora.ru/corpora-sem.html). Процедура аннотации корпуса основана на семантическом словаре, в котором каждое словарное значение слова имеет свой словарный вход[2] и представляется в виде набора параметров, ср.:

[2] С точки зрения представления семантической информации в корпусе традиционная омонимия (*жить в МИРЕ и согласии ~ МИР животных*) и многозначность (*ВСПЫШКА гнева ~ фотографическая ВСПЫШКА*) считаются явлениями одного порядка; соответственно,

ПРИПЛЫТЬ – "глагол движения", "некаузативный глагол", "приставочный глагол".

Фасетность, т. е. параметризация словарного значения по нескольким основаниям, не ведет к семантической омонимии: последняя возникает, если разные значения слова относятся к разным лексико-семантическим классам:

ЗОЛОТОЙ 1 (*золотое кольцо*) – "относительное прилагательное";
ЗОЛОТОЙ 2 (*золотые кудри*) – "прилагательное цвета", "относительное прилагательное";
ЗОЛОТОЙ 3 (*золотой ребенок*) – "прилагательное оценки", "качественное прилагательное".

Программа первичного семантического парсинга работает без учета контекста и приписывает лемме семантические признаки, относящиеся ко всем ее значениям. Естественно, это создает много шума при поиске. Во-первых, некоторые слова вообще не получает семантического разбора, если они отсутствуют в семантическом словаре. Во-вторых, часть слов получает несколько альтернативных разборов: это слова с семантической омонимией. В третьих, слова могут иметь ошибочный разбор (при разборе "вручную" эти слова получили бы разбор, отличный от словарного). Не будем также забывать, что в случае, если семантическая разметка накладывается на корпус с неснятой морфологической омонимией, процент ошибок многократно умножается.

В нашей предыдущей работе Кобрицов, Ляшевская 2004 обсуждалась проблема снятия семантической омонимии с помощью глубинных фильтров - на основе глобальных правил сочетаемости семантических классов, например, "названия одежды не могут употребляться в роли субъекта при глаголах эмоции", ср. *амазонка* <‘человек’, ~~‘одежда’~~> *рассмеялась*. Как показывает практика, точность глобальных правил далека от 100 процентов (Кобрицов 2004).

Поверхностные фильтры, которые на входе содержат комбинацию из 2-х, 3-х и т. д. лексем (или даже словоформ), ср.:

термин "разрешение семантической омонимии" (word-sense disambiguation) охватывает оба явления.

w1 *до*
w2 *сих*
сей (A-PRO) <"указательное мест.">
w3 *пор*
по́ра (S) <"предметное имя", "простр.:отверстие">
по́ра (S) <"предметное имя", "простр.:пустота">
пора́ (S) <"непредметное имя", "период времени">
пора́ (PRAEDIC)

→ пора́ (S) <"непредметное имя", "период времени">,

напротив, почти всегда позволяют предсказать единственно правильный разбор. Проблема лишь в том, что для того чтобы существенно уменьшить семантическую омонимию в корпусе, требуются тысячи таких фильтров. Таким образом, перед нами стояла задача максимально автоматизировать работу по подготовке исходного материала для экспертов и выделить самые частотные коллокации для создания наиболее эффективных поверхностных фильтров.

3.2. Частотные устойчивые коллокации

Списки устойчивых коллокаций были получены на базе корпуса публицистики[3], включающего материалы московских и региональных газет за 1998-2004 гг., новостей, радиоинтервью, а также мемуарную литературу. Объем обработанного корпуса составил 15,8 млн слов. Были получены реестры двусловных и трехсловных "жестких" коллокаций, в которых составляющие непосредственно примыкают друг к другу (т. е. расстояние между словами не превышает единицы). Наш алгоритм не учитывал пары и тройки словоформ, разделенные границами предложения, а также любыми знаками препинания. Таким образом, сюда попали сочетания типа *Московский комсомолец* и не попали сочетания типа *газета "Московский комсомолец"*. Словосочетания с переменой мест составляющих считались разными коллокациями (ср. *российское государство* и *государство российское*).

Реестры коллокаций были обработаны с помощью лингвистических и статистических методов. На основании информации о частях речи из списков были исключены коллокации, не образующие

[3] Входит в состав Корпуса современного русского языка.

синтаксического единства, типа "и в" (CONJ + PR), "в самом" (PR + APRO) и др. (метод Джастесона и Каца [Justeson, Katz 1995]). Был также использован стоп-лист малоинформативных слов, например *или*/*этот*/*мой* + S.

Как известно, ни один из существующих методов статистического ранжирования коллокаций (см. их обзор в [Manning, Schütze 1999; Pearce 2002; Jiangsheng Yu et al. 2003]) не позволяет с уверенностью различить "хорошие" и "плохие", т. е. случайные, коллокации. В список представляемых на суд эксперта оборотов были включены все коллокации, абсолютная частотность которых превышала 100 употреблений. Оставшиеся коллокации были ранжированы с помощью формулы

$$MI = \frac{frequency(w1,w2)^2}{frequency(w1) \bullet frequency(w2)},$$

использовавшейся ранее при обработке Кембриджского корпуса английского языка. Верхняя часть полученного списка была также включена в short-list.

Наконец, была учтена информация о дискурсивных сдвигах значения односложных элементов (ср. существительное в функции предлога *типа*) и о более чем 3-словных оборотах, полученная из работ [Русская грамматика 1980; Зализняк 1977/2003; Рогожникова 2003; Шведова 1960/2003][4].

3.3. Характеристика поверхностных фильтров

Поверхностные фильтры, работающие на базе устойчивых коллокаций, включают данные (1) о лемме, (2) о частеречных, (3) словоклассифицирующих и (4) словоизменительных признаках составляющих, (5) об их исходной семантической разметке, а также (6) о некоторых грамматических и лексико-семантических характеристиках ближайшего контекста (например, "родительный падеж" для оборота *типа кого-чего-л.*).

Для увеличения скорости обработки больших массивов оборотов эксперт может сортировать список по любому из параметров,

---

[4] Благодарим также С.А. Шарова, предоставившего нам списки имен и фамилий людей, полученных при анализе его собственного корпуса публицистики – они также были добавлены в БД коллокаций.

например, обрабатывать все обороты с вариантом семантического разбора "период времени" или все 2-грамы с наиболее частотным правым членом (*А-ая страна, А-ый вопрос*). Кроме того, эксперт может проследить дерево вложенных коллокаций, ср.:

НЕ ГОВОРЯ О <предл. пад.>
НЕ ГОВОРЯ О том, что...
НЕ ГОВОРЯ О том, как...
я НЕ ГОВОРЮ О <предл. пад.>
мы НЕ ГОВОРИМ уже О том что...

Говоря о частотных коллокациях, обычно имеют в виду, что частотность употребления двух слов рядом друг с другом ведет к тому, что в их значении появляются нетривиальные компоненты, нарушающие строгий принцип композициональности значения [Manning, Schütze 1999][5]. В связи с этим на выходе поверхностных фильтров может находиться не только один семантический разбор из нескольких данных, но и новый, несловарный разбор. (Заметим, что на этапе подготовки short-листа коллокаций в список были включены обороты, у которых межлексемная и семантическая омонимия изначально отсутствовала).

С помощью поверхностных фильтров в семантическую разметку также добавляется особая информация о функционировании слова в составе оборота. В частности, таким образом размечаются:

– сложные служебные лексические единицы: составные предлоги, союзы, наречия, частицы, вводные слова (*в связи с, в случае если, на самом деле* и др.);
– топонимы (*Нижний Новгород*);
– неоднословные обозначения лиц (*Владимир Путин, В. Путин, президент Путин*) и др.

Поверхностные фильтры позволяют зафиксировать дискурсивные сдвиги частеречной принадлежности: существительное в функции предлога (*типа, вида*); междометие или наречие в функции

[5] Классические идиомы, впрочем, оказались большей частью вне зоны нашего внимания – их частотность в текстах слишком мала. В то же время были обработаны высокочастотные коллокации, в которых не отмечается диффузии значения, ср. *министр финансов, президент Путин*.

существительного (*пройти на ура; перенести на завтра*); субстантивацию (*в открытую*).

Отмечаются следующие случаи выветривания значения:

– употребление глагола в качестве лексической функции (ср. *вступить в силу*);
– употребление прилагательного в качестве лексической функции Magn (*круглый дурак*);
– включение предлога в модель управления глаголов и имен (*делить на…, проблема с…, министр по…, один из…*);
– идиоматическое употребление одной или нескольких составляющих (ср. *круглый стол*: весь оборот относится к классу "мероприятие").

4. Выводы и обсуждение результатов

4.1. Оценка эффективности работы фильтров

Подведем итоги, достигнутые после создания фильтров на основе 4500 частотных устойчивых коллокаций.

Оценка эффективности проводилась в три этапа. На первом этапе было подсчитано количество словоупотреблений в исходном корпусе и число нераспознанных словоупотреблений. Уровень распознавания текста (лемматизации) составил приблизительно 98%. Некоторые такие слова (*госдума*, *СМИ* и др.) были обнаружены в списке высокочастотных коллокаций. После определения их исходной формы и частеречной принадлежности и применения фильтров к текстам точность разметки была повышена на 0,26%.

На втором этапе было подсчитано число словоформ с межлексемной омонимией (в среднем 1,53 разборов на каждое словоупотребление). Разрешение межлексемной омонимии с помощью поверхностных фильтров позволяет достичь точности разметки 1,13 разборов на одно словоупотребление.

На последнем этапе оценивалось качество собственно семантической разметки. Число слов, размеченных по лексико-семантическим параметрам, составило в настоящей версии корпуса 67% всех словоупотреблений. Однако точность семантической разметки представляется разумным рассчитывать без учета слов,

относящихся к предлогам, союзам, частицам и другим частям речи, на которые семантическая разметка не распространяется. По данным подкорпуса со снятой омонимией, к знаменательным частям речи должно принадлежать порядка 75,9% всех словоупотреблений. Таким образом, относительно них доля семантически размеченных слов составила 85,7% (13,5 млн из 15,8 млн) .

Мы лишены возможности сравнить наш корпус с некоторым "золотым стандартом", ибо русских корпусов, размеченных "вручную" описанным выше методом, не существует. Это, безусловно, затрудняет оценку правильности семантической разметки.

Предлагается два показателя оценки относительного улучшения качества разметки. Первый – показатель полного снятия омонимии,

$$WSD = \frac{x_f - x_0}{N} \cdot 100\%,$$

где x_f – число слов с единственным правильным разбором в корпусе после применения фильтров, x_0 – аналогичное число в исходном корпусе, N = 13 538 782 – база для оценки точности разметки, общее число слов знаменательных частей речи.

Коэффициент WSD учитывает те случаи, когда фильтр однозначно снимает семантическую омонимию:

$$w\{s_1, \ldots s_n\} \rightarrow w\{s_i\}.$$

Для тех случаев, когда фильтр снимает только часть омонимии, например,

$$w\{s_1, \ldots s_n\} \rightarrow w\{s_1, s_2\} -$$

используется коэффициент

$$WSR = \left(\frac{s(N-x_0)}{N-x_0} - \frac{s(N-x_f)}{N-x_f} \right) \cdot 100\%,$$

где $N-x_0$, $N-x_f$ – число слов с неснятой омонимией, а $s(N-x_0)$, $s(N-x_f)$ – общее количество разборов у слов с неснятой омонимией.

Предварительные результаты работы показали, что система, построенная на 1000 наиболее частотных коллокаций, позволила

разрешить неоднозначность в исходном корпусе для 800 тыс. словоупотреблений полностью (WSD ≈ 6%) и для 100 тыс. словоупотреблений частично (WSR ≈ 1,5%).

Эксперименты на финальной стадии проекта показали, что фильтры, построенные на 4,5 тысячах частотных коллокаций, позволяют уменьшить семантическую неоднозначность для разных тестовых подкорпусов в объеме от 3,3% до 6% от всей многозначности в корпусе.

4.2. Другие полезные результаты и перспективы.

В ходе работы над проектом были достигнуты некоторые побочные результаты, которые позволили улучшить качество разметки в Национальном корпусе русского языка. В первую очередь, БД устойчивых коллокаций дала возможность аннотировать и ввести в корпус информацию о неоднословных лексических единицах. Помимо этого, был проведен анализ ошибочно аннотированных элементов текста. В результате был пополнен словарь Корпуса; внесены исправления в семантический словарь (изменена структура значений и семантические признаки некоторых слов); пополнен словарь аббревиатур (ср. частотные коллокации "см. рис.", "под ред.").

Мы предполагаем, что база данных устойчивых коллокаций может быть использована для создания более эффективных процедур построения глубинных фильтров. Мы собираемся проанализировать наиболее типичные сочетания семантических признаков в поверхностных фильтрах, и, возможно, некоторые такие случаи будут обобщены как глубинные фильтры. Интересен также анализ отрицательного материала – выявление конфликтующих семантических признаков. Особую задачу представляет исследование списков частотных коллокаций, отвечающих критериям статистической релевантности, но которые эксперты тем не менее не признали лингвистически релевантными.

5. Литература

Большаков И.А., Галисия-Аро С.Н. Сколько страниц на данном языке содержит Интернет? // Труды международной конференции Диалог'2003. М., 2003.

Борисова Е.Г. Коллокации. Что это такое и как их изучать? М., 1995.

Зализняк А.А. Грамматический словарь русского языка. М., 1977. 4-е изд.: М., 2003.

Добровольский Д.О. Корпус параллельных текстов как инструмент анализа литературного перевода. Труды международной конференции Диалог'2003. М., 2003.

Кобрицов Б.П. Методы снятия семантической многозначности // Научно-техническая информация, сер.2, 2004а, N 2.

Кобрицов Б.П. Модели многозначности русской предметной лексики: глобальные и локальные правила разрешения омонимии. Автореф… канд. филол. наук. М.: РГГУ, 2004б.

Кобрицов Б.П., Ляшевская О.Н. Автоматическое разрешение семантической неоднозначности в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. Под ред. И.М.Кобозевой, А.С.Нариньяни, В.П.Селегея. М., 2004.

Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2005. М., 2005.

Копотев М. «Несмотря на» «потому что», или Многокомпонентные единицы в аннотированном корпусе русских текстов. Диалог'2004. М., 2004.

Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Национальный корпус русского языка как инструмент семантико-грамматического исследования лексики // Международная конференция "Корпусная лингвистика - 2004". Тезисы докладов. СПб.: СПбГУ, 2004. С. 50-51.

Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Опыт семантического расширения морфологической разметки: таксономическая классификация лексики в Национальном корпусе

русского языка // Научная и техническая информация, сер. 2. Информационные процессы и системы (в печати).
Русская грамматика. М., 1980.

Рогожникова Р. П. Словарь эквивалентов слова. М., 2003.

Шведова Н.Ю. Очерки по синтаксису русской разговорной речи. М., 1960. 2-е изд.: М., 2003.

Шмелев Д.Н. О семантических изменениях в современном русском языке // Шмелев Д.Н. Избранные труды по русскому языку. М., 2002.

Black E. An experiment in computational discrimination of English word senses, in IBM Journal, 32(2), 1988. P. 185-194.

Bolshakov I.A. Stable coordinate pairs as a specific resource of language // Apresjan Ju. D., Iomdin L.L. (eds.) East – West Encounter: Second International Conference on Meaning ⇔ Text Theory. Moscow, 2005.

Dagan I., Itai A., Schwall U. Two languages are more informative than one // Proceedings of the ACL, 1991 (29). P. 130-137.

Demetriou G.C. Lexical disambiguation using constraint handling in Prolog (CHIP) // Proceedings of the European Chapter of the ACL, 1993 (6). P. 431-436.

Demetriou G. and Atwell E. A domain-independent semantic tagger for the study of meaning associations in English text // Bunt H., van der Sluis I., Thijsse E. (eds.), Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4), Tilburg, Netherlands, 2001. P. 67-80.

Gale W.A., Church K.W., Yarowsky D. A method for disambiguating word senses in a large corpus // Computers and the Humanities, 1992, 26. P. 415-439.

Guthrie J.A., Guthrie L., Wilks Y., Aidinejad H. Subject-dependent co-occurrence and word sense disambiguation // Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA., 1991. P. 146-152.

Hearst M.A. Noun homograph disambiguation using local context in large text corpora // Proceedings of the 7th conference, UW Centre for the New OED & Text Research Using Corpora, 1991.

Hirst G. Semantic interpretation and the resolution of ambiguity, in Cambridge, 1986.

Ide N., Veronis J. Introduction to the special issue on word-sense disambiguation: the state of the art // Computational Linguistics, 24:1. P. 1-14.

Jiangsheng Yu, Zhihui Jin, Zhenshan Wen. Automatic Detection of Collocation // The 4th Chinese lexical semantics workshop, Hong-Cong, 2003. http://icl.pku.edu.cn/yujs/papers/pdf/col.pdf.

Justeson J.S., Katz S.M. Technical terminology: some linguistic properties and an algorithm for identification in text // Natural Language Engineering, 1995, 1(1). P. 9-27.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the 11th EURALEX International Congress. Lorient, France : Universite de Bretagne-Sud, 2004. P. 105-116.

Lesk M. “They said true things, but called them by wrong names” – vocabulary problems in retrieval systems // Proceedings of the 4th Annual Conference of the University of Waterloo Centre for the New OED, 1988.

LEXICOM 2005. Course handbook of the 5th Annual Workshop in Lexicography and Lexical Computing. Brno, 2005.

Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing // Cambridge, Massachusetts: The MIT Press, 1999. Ch. 5. Collocations. http://nlp.stanford.edu/fsnlp/promo/colloc.pdf.

Masterman M. The thesaurus in syntax and semantics // Mechanical Translation, 4, 1957. P. 71-72.

Patrick A. B. An exploration of abstract thesaurus instantiation. M. Sc. thesis, University of Kansas, Lawrence, Kansas, 1985.

Pedersen T., Banerjee S., Patwardhan S. Maximizing Semantic Relatedness to Perform Word Sense Disambiguation // University of Minnesota Supercomputing Institute Research Report UMSI 2005/25, March 2005. http://www.msi.umn.edu/general/Reports/rptfiles/2005-25.pdf.

Pearce D. A comparative evaluation of collocation Extraction Techniques // Third International Conference on Language Resources and Evaluation. May, 2002. Las Palmas, Canary Islands, Spain. 2002. http://www.informatics.susx.ac.uk/users/darrenp/academic/dphil/publications/data/Conferences/lrec2002/paper.pdf.

Stevenson M., Wilks Y. Large vocabulary word-sense disambiguation // Ravin Y., Leacock C. (eds.) Polysemy: Theoretical and Computational Approaches. Oxford, 2002. P. 161-177.

Stevenson M. Word Sense Disambiguation: The Case for Combining Knowledge Sources. CSLI Publications, Stanford, CA, 2003.

Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network // Proceedings of the International Conference on Information & Knowledge Management (CIKM), 2, 1993. P. 67-74.

Weiss S. Learning to disambiguate // Information Storage and Retrieval, v.9, 1973.

Wilks Y., Fass D., Guo C., Mcdonald J.E., Plate T., Slator B.M. Providing Machine Tractable Dictionary Tools // Machine Translation, 5, 1990. P. 99-154.

Yarowsky D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora // Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, 23-28 August, Nantes, France, 1992. P. 454-460.

Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods // Proceedings of the ACL'1995, 33.

Word-sense disambiguation in mass media texts: shallow rules and statistic evaluation

Boris P. Kobritsov, Olga N. Lashevskaja, Olga Ju. Shemanaeva

This report presents a method of word sense disambiguation [Hirst 1986; Ide, Veronis 2002; Stevenson 2003 и др.] that uses shallow rules, or rigid patterns (Weiss 1973). These rules provide the highest degree of accuracy but at the same time they are considered to be most labour-concuming. We explore the method of automatic compiling of the high-frequency stable collocations database combined with its subsequent (half)manual annotating. The corpus of mass media text serves as a source of our investigation, because the portion of stock phrases as *vstupit' v silu* 'join into force', *uroven' zhizni* 'standard of living' is particularly great in news, newspapers and journals. At present, the system of 4500 shallow rules is constructed. Our experiments on applying shallow rules to various subcorpora show that semantic ambiguity is reduced from 3,3 до 6%.