

Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей

Кобрицов Б.П.
ВИНИТИ РАН
neuralman@yandex.ru

Ляшевская О.Н.
ВИНИТИ РАН
olesar@mail.ru

Толдова С.Ю.
МГУ им. М.В.Ломоносова
toldova@pisem.net

Аннотация

В работе описывается серия экспериментов по снятию семантической неоднозначности глаголов с использованием информации об их моделях управления, извлеченной из толковых словарей, а также из специализированного словаря глагольного управления. Авторы приходят к выводу о том, что информация, извлекаемая автоматически из словарей, не позволяет кардинальным образом понизить степень многозначности. Специализированный словарь может служить в качестве исходной базы для дальнейшего обучения по корпусу. Эксперимент показал, что разные признаки моделей управления в разной степени влияют на возможности сокращения числа значений, связанных с глаголом в некотором контексте. Наибольшей различительной силой обладают периферийные (факультативные) актанты, а также такая семантическая характеристика актантов как абстрактность. Словарная информация позволяет выделить класс глаголов, для которых модель управления служит надежным признаком для разрешения многозначности. В общем случае даже после обучения на корпусе информация о грамматических и обобщенных семантических свойствах существительных, входящих в модель управления глагола, не позволяет полностью разрешать многозначность. Она дает возможность разбить множество всех значений на классы, отсеять периферийные значения и, тем самым, в значительной мере снизить степень многозначности. Однако радикальное понижение степени многозначности можно достичь только с учетом семантических характеристик актантов, а для ряда глаголов только с привлечением информации о более узком тезаурусном классе актантов.

1. Введение

1.1 Постановка задачи

Разрешение семантической неоднозначности является одной из наиболее сложных проблем в системах автоматической обработки естественного языка, в том числе в системах семантической разметки корпуса текстов.

Особенно сложной и важной задачей представляется снятие неоднозначности для глаголов. Глаголы представляют собой грамматический класс лексем с наиболее развитой системой полисемии. В то же время они служат синтаксическим и семантическим ядром предложения. В исследовании мы исходили из общепринятого положения о том, что базовые свойства глагола определяются их моделями управления (см., например, [15:120]).

Выделение моделей управления (МУ) во многих

системах NLP либо базируется только на частотных характеристиках контекста, что снижает его точность, либо является трудоемким ручным процессом, либо использует специальные лексикографические ресурсы, такие, например, как WordNet [3], FrameNet [6] и др., создание которых также требует больших трудозатрат. Для русского языка на данный момент такая система недоступна. Встает вопрос, в какой степени можно использовать доступные компьютерные источники, такие как толковые словари, для извлечения информации о значениях и соответствующих им моделях управления глаголов.

Таким образом, целью исследования являлась оценка возможности метода, базирующегося на установлении соответствия между множеством МУ глагола и множеством его возможных значений. Метод предполагал извлечение информации о связи МУ с конкретным значением глагола из лексикографического источника и из морфологически размеченного корпуса, выделение необходимых диагностических параметров МУ, наложение выделенных моделей на тестовый корпус. При обучении на корпусе учитывались как морфо-синтаксические свойства именных групп, так и их семантические свойства в терминах обобщенных таксономических классов (одушевленность, абстрактность/конкретность).

1.2 Предыдущие работы

Современные системы автоматического разрешения лексико-семантической многозначности (word sense disambiguation), исходя из тезиса о том, что конкретное значение лексемы связано с определенным типом контекста, в котором данное значение встречается. Задача “типизации” контекста решается либо чисто статистическими методами [4], либо с привлечением экспертных ресурсов. В качестве источника признаков контекста используют данные лексикографических источников (см., например, [8]), корпусные данные (статистические характеристики контекста в корпусе), данные параллельных текстов (ср. принцип «один смысл – один перевод» в [2]).

Что касается статистических методов, то для разрешения многозначности используются как контролируемые методы обучения, так и неконтролируемые ([1], [4], [8] и др., см. также обзоры в [10] и [17]). Большинство таких систем

базируются на байесовской модели или на модели канала с шумом. В работе [4] сообщается о достижении 90% точности для шести существительных с достаточно четко различимыми смыслами. Данный метод активно разрабатывался на материале английского языка. Он требует большого корпуса, размеченного вручную. Потенциальными признаками контекста являются все лексемы из достаточно большого окна. При неконтролируемом обучении (см., например [10]) невозможна семантическая разметка с приписыванием тому или иному слову семантического тэга, задача сводится к классификации множества контекстов на группы и их различению (discrimination).

Для задач данного проекта было важно учесть опыт второго подхода, при котором используются специализированные лексикографические ресурсы. Такие методы предполагают либо первичную полуавтоматическую разметку тренировочного корпуса (ср. проект Senseval [11], либо использование тезаурусов и словарных систем, таких как Wordnet, FrameNet, VerbNet. Технологии применения данных систем активно разрабатываются в проектах по семантическому аннотированию корпусов на многих языках, в том числе при разрешении многозначности глаголов с использованием моделей управления: [6], [7], [12], [13], [14].

Для русского языка лексикографические системы такого типа находятся в стадии разработки (RusNet [15]). В рамках проекта RusNet проводились пилотные эксперименты по применению лексикографических источников для извлечения моделей управления (см. [20]). Однако данный проект предполагает задачу лексикографического описания глагола, а не снятие омонимии в корпусе.

2. Идея исследования

2.1 Семантическая разметка корпуса и проблема семантической неоднозначности

Проводимый нами эксперимент должен был помочь найти решение более общей задачи, возникшей при работе над созданием таксономической (семантической) разметки в Национальном корпусе русского языка (НКРЯ, подробнее о корпусе и таксономической разметке см. [19], [21]). Если для морфологической разметки система категорий, в терминах которой происходит разметка, достаточно универсальна, то для семантической разметки не существует такого единого перечня семантических ярлыков. Системы автоматического или полуавтоматического разрешения неоднозначности существенным образом различаются «глубиной» различения многозначности, выбором словаря или лексической классификации, к которой привязана семантическая аннотация и т.п. Параметры семантической разметки прежде всего зависят от того, каковы потребности пользователей конечного продукта и каким способом (и с какими затратами)

разработчики собираются добиться нужного результата.

Для семантической разметки НКРЯ был выбран принцип небольшого количества таксономических классов для каждой из частей речи. Это объясняется рядом практических соображений. Во-первых, «прямой» поиск, без построения дерева вложенных подклассов, обеспечивает быструю выдачу результатов. Во-вторых, ситуация, когда все названия семантических классов обозримы, видны в одном окне компьютера, помогает пользователю быстрее сориентироваться в системе классификации и, соответственно, быстро задать поисковый запрос.

В настоящее время перед разработчиками корпуса стоит задача повышения точности разметки и снижения уровня «шума» в результатах поиска. В нашем проекте она связана с учетом разных значений многозначных и омонимичных слов и с уменьшением семантических тэгов, связанных с лексемой в конкретном контексте. При этом разрабатываемые методы должны быть относительно просты, не требовать ручной разметки большого массива данных. Мы также не можем использовать результаты полного синтаксического разбора, поскольку синтаксическая разметка в корпусе отсутствует.

В качестве основного метода при разрешении неоднозначности в корпусе используется метод глубинных и поверхностных фильтров. Фильтры учитывают лексико-грамматическую сочетаемость слов. Для каждой леммы строится своя система фильтров. Технология фильтров активно применяется при разрешении неоднозначности для существительных. Однако такая технология предполагает создание тысяч фильтров и требует работы эксперта с каждой лексемой отдельно. В силу этого разработка технологий, позволяющих создавать хотя бы часть фильтров в автоматическом или полуавтоматическом режиме для всего грамматического класса, а также достаточно просто уменьшать количество тэгов, связанных с некоторой лексемой в контексте является, актуальной.

Для глагольных лексем основная сочетаемость определяется способностью глаголов присоединять тот или иной набор актантов (существительных, существительных с предлогами, придаточных, инфинитивных оборотов и т.п.). Утверждение о том, что семантический сдвиг в глаголе сопровождается изменением свойств его актантов, является основой для многих семантических описаний глагольных значений.

Таким образом, в основе нашего исследования лежала гипотеза о том, что использование модели управления глагола, включающей морфологические характеристики актантов и семантические характеристики именных актантов, позволяет уменьшить количество тэгов при семантической разметке глаголов. Действительно, большинство систем, позволяющих извлекать актантную

структуру глагола из текста, используют специально разработанные достаточно подробные таксономические классификации существительных. Однако, создание такого ресурса также отдельный трудоемкий процесс. К тому же, существительные имеют высокую степень многозначности и эффективное использование их более точных семантических характеристик предполагает предварительное снятие омонимии для существительных. Возникают следующие вопросы, возможно ли понизить уровень многозначности, если учитывать только самые общие семантические характеристики и в какой степени информация об основных актантах глагола может быть извлечена из существующих словарей.

Эксперимент должен был ответить на следующие вопросы:

- в какой степени можно использовать информацию автоматически или полуавтоматически извлеченную из лексикографических источников;
- в какой степени данные о МУ глагола с использованием минимальной информации о семантическом классе актантов (одушевленность vs. неодушевленность, абстрактность vs. конкретность) позволяет понизить степень многозначности;
- каковы возможные методы уточнения МУ по обучающему корпусу;
- каковы возможные методы расширения МУ семантическими признаками актантов.

2.2 Актуальность исследования

Если для английского и китайского языков с широким развитием проектов по ручной аннотации моделей управления в корпусах большого объема (FrameNet, PropBank, NomBank, Chinese PropBank), подход, основанный на статистической тренировке, можно считать принципиально реализованным, то для русского языка таких открытых ресурсов на данный момент не существует. Наш проект учитывает опыт по использованию моделей управления, извлеченных из модельного словаря, в том числе и эксперименты, проводимые в группе разработчиков RusNet под руководством И.В. Азаровой. В нашем эксперименте выбраны несколько лексикографических источников. Мы учитывали только информацию об обобщенных семантических признаках существительных. Если для проекта RusNet главная задача «добыча» подробной информации о МУ и значениях глагола для создания лексикографического ресурса, то наша задача оценить, что могут дать сравнительно дешевые методы извлечения информации из доступных на данный момент источников без использования сложного синтаксического модуля и с учетом минимальной информации о семантических признаках существительных.

3. Описание методов, алгоритмов, экспериментов.

3.1 Основные используемые понятия

В рамках проекта под многозначностью понимался термин, используемый в прикладных системах. В эксперименте рассматривались все возможные значения, которые могут быть приписаны одному и тому же графическому входу словаря. Каждому из исследуемых глаголов (ключевых глаголов) было сопоставлено множество семантических тэгов, выделяемых в соответствующем источнике (словаре или наборе тэгов корпуса). Каждому из значений (каждому тэгу) ставилось в соответствие множество словарных МУ, извлеченных из словарей и в результате обучения по корпусу. Каждому из предложений обучающего и тестируемого корпусов с соответствующим ключевым глаголом также ставилась в соответствие некоторая МУ. В последнем случае говорим о контекстной (реализованной в предложении) МУ.

МУ представляет собой множество позиций контекста (актантов), связываемых с данным значением глагола. Актант характеризуется своими грамматическими признаками. Чаще всего актантами являются существительные и местоимения. Они характеризуются предложно-падежной формой, в которой регулярно выступают в контекстах. Например, одна из МУ для глагола *найти* (*Олег нашел кошелек в сквере*) выглядит следующим образом {S&nom; S&acc; v+S&loc}, где S – существительное или местоименное существительное, а nom, acc и loc – именительный, винительный и предложный падежи. Кроме того, учитываются актанты, оформленные инфинитивным оборотом или придаточным предложением. Для таких актантов указывается либо форма глагола (инфинитив), либо некоторые диагностические признаки конструкции (союз, с помощью которого присоединяется придаточное и т.п.). Например, одна из моделей управления, связанная с глаголом *найти* (как в *Вы не находите, что он прав*), выглядит следующим образом: {S&nom, что}. Здесь и далее используются также следующие обозначения: gen – родительный, dat – дательный, ins – творительный. Кроме того, в характеристики актантов могут входить ограничения на семантический класс актанта, такие как одушевленность (anim/inan), конкретность / абстрактность (concr/abstr).

Таким образом, для каждого глагола, отобранного для эксперимента (ключевого глагола (V_0)), использовались следующие типы данных:

- набор семантических тэгов, связанных с глаголом в соответствующем источнике $S(V_0, источник_i) = \{s_1, s_2, \dots, s_n\}$;
- набор словарных моделей управления $MU(V_0, источник_i) = \{MU_1, \dots, MU_k\}$;
- множество классов значений, полученных в результате связывания некоторой МУ со

множеством соответствующих этой МУ значений: например, МУ {S v+S loc} для глагола *идти* связана со следующим множеством корпусных значений {move, move_metaph, LF} (количество значений в данном классе - 3).

После этапа обучения мы получали множество уточненных МУ. После добавления семантической информации – МУ с семантическим расширением.

Если считать, что первоначально семантический разметчик ставит в соответствие каждому предложению с ключевым глаголом множество всех значений по источнику, т.е. $S_i(V_0) \rightarrow \{s_1 \dots s_n\}$, то задачей эксперимента было сократить число значений в данном множестве, в идеале до одного.

3.2 Исходные данные.

В качестве словарей, из которых извлекалась информация, были выбраны Малый академический словарь (МАС), словарь Ожегова (ОЖ), а также специализированный словарь Апресян-Палл [17] (АПР).

В качестве обучающего корпуса использовался подкорпус НКРЯ объемом в 4,5 миллионов словоупотреблений со снятой морфологической омонимией. В состав корпуса входила как художественная литература, так и журнальные и газетные статьи. В данном подкорпусе каждому словоупотреблению приписан единственный полный морфологический разбор, а также множество семантических тэгов. Для выделения МУ мы использовали эту морфологическую информацию, включая информацию об одушевленности, а также информацию об абстрактности vs. конкретности соответствующих существительных. Эта информация извлекалась из первого семантического тэга для первого значения существительного. В качестве одного из источников для формирования множества значений использовалась таксономическая (семантическая) разметка корпуса.

Для эксперимента были отобраны 400 высокочастотных глаголов из разных семантических классов. Для более подробного анализа результатов было отобрано 10 глаголов из данного списка.

3.3 Особенности семантической разметки в корпусе

В корпусе существует таксономическая (семантическая) разметка, т.е. каждое значение слова задается набором семантических ярлыков, свидетельствующих о принадлежности лексемы к тому или иному таксономическому классу, например:

найти

- 1) 'каузация обладания' (*найти кошелек*);
- 2) 'восприятие' (*найти его больным*);
- 3) 'движение' (*коса нашла на камень*)

Как видно из примера, многозначной лексеме первоначально приписывается все множество возможных ярлыков.

Как уже указывалось, система семантических тэгов в НКРЯ строится на сравнительно небольшом количестве семантических классов и не учитывает дробную классификацию значений. В эксперименте использовалась расширенная система помет. Было введено несколько дополнительных семантических классов. В общей сложности число классов не превышало пятидесяти. В состав тэгов также вошли пометы, обозначающие метафорический переход по некоторому значению. Разница в нюансах такого метафорического переноса не учитывалась: например, для глагола *идти*, метафорическим переносом считается его употребление в предложении: *Чай шел с Востока* и употребление в предложении *Армия шла от победы к победе*¹. Также в большинстве случаев глагол получал единую помету для всех контекстов, в которых он был употреблен в качестве лексической функции (LF). Таким образом, число корпусных значений для одного глагола не превышало десяти.

3.4 Методы оценки результатов эксперимента

Основной упор при анализе результатов делался на содержательный анализ. Что касается числовых оценок, то результаты применения метода к каждому глаголу оценивались отдельно, общая оценка строилась как среднее по тестируемым глаголам. Мы использовали следующие количественные характеристики:

- 1) покрытие тестового корпуса МУ источника, т.е. какому проценту предложений из корпуса возможно приписать одну из моделей, указанных в источнике;
- 2) максимальная степень многозначности глагола: количество семантических тэгов в источнике ($S_{\max}(W)$);
- 3) средняя степень многозначности для глагола: среднее количество семантических тэгов, приписанных глаголу в контексте: $S_{cp}(W) = s_f(N)/N$, где $s_f(N)$ – результирующее количество тэгов, приписанных глаголу в выборке, N – количество предложений с данным глаголом в выборке;
- 4) показатель полного разрешения неоднозначности:
- 5) $WSD = (x_1/N) * 100\%$, где x_1 – число глаголов с единственным семантическим тэгом после применения метода;
- 6) показатель понижения степени неоднозначности: $WSR = ((S_{\max}(N) - s_f(N)) / ((S_{\max}(N)/N - 1) * 100)\%$, где S_{\max} – первоначальное общее количество тэгов, s_f – результирующее количество тэгов;
- 7) точность алгоритма: число примеров, где правильный семантический тэг остался в числе множества семантических тэгов, приписанных данному глаголу в соответствующем примере, поделенное на общее число примеров.

¹ Отметим, что эти 2 метафорических переноса можно различить, если учесть семантический класс локативных актантов (конкретный в первом случае и абстрактный во втором), но в силу общего принципа избегать излишне дробной классификации значений, мы остановились на решении их не разделять.

Число значений, выделяемых для разных глаголов в словарях, колеблется от 2 до 25. Результат применения метода к глаголу сильно зависит от числа значений в источнике. То есть, если первоначальное количество значений 2, то результирующее значение степени многозначности обычно не превышает 1,2 тэга на глагол, если значений 10, то не меньше, чем 2,5-3 тэга на глагол. В последнем случае уменьшение количества тэгов тоже можно считать положительным результатом, даже, если при этом степень многозначности остается достаточно высокой. В силу этого мы ввели для оценки характеристику (6). Мы исходили из следующего: при полном разрешении многозначности наша задача уменьшить первоначальное (словарное) количество тэгов, связанных с глаголом, до одного тэга на глагол. Такое понижение степени многозначности соответствует 100% успешности работы алгоритма.

Необходимо отметить, что реальная степень многозначности в размеченном вручную корпусе – более 1 значения на глагол. При ручной разметке обучающего корпуса предложению приписывалось значение, которое эксперт был в состоянии приписать, не обращаясь к контексту за пределами анализируемого предложения. В некоторых предложениях не был выражен ни один актанта (подробнее см. п.3.11). Таким предложениям приписывалось несколько тэгов. Однако при оценке результатов мы не учитывали это обстоятельство.

Характеристики (6) и (7) связаны между собой. Если степень многозначности понизилась на 0%, т.е. ни один тэг не был удален, то точность должна быть близка к 100% (за исключением тех случаев, где предложению не может быть приписан ни один семантический тэг из множества тэгов источника, например, в АПР для глагола *найти* отсутствует значение 'движение', как в *Нашла коса на камень*). Чем больше тэгов вычеркивается, тем больше вероятность ошибки.

3.5 Основные этапы

Условно исследование можно разбить на 4 этапа. Первый, подготовительный этап, включал подготовку исходных данных для проведения эксперимента. На данном этапе решались следующие задачи:

- 1) извлечение информации о модели управления и значении русских глаголов из электронных версий лексикографических источников (МАС, ОЖ, АПР);
- 2) выборка эталонных глаголов из числа наиболее частотных, удовлетворяющих требованиям многозначности и разнообразия глагольного управления;
- 3) создание обучающего и тестового корпусов.

Второй этап также относился к процедуре подготовки данных и включал решение следующих задач:

- 4) извлечение контекстных МУ, соответствующих словарным моделям управления;
- 5) выделение "триад": базовая модель управления -

значение по лексикографическим источникам - семантический класс по классификации Национального корпуса русского языка.

Надо отметить, что этапы 4) и 5) потребовали гораздо больших усилий, чем планировалось заранее. Дело в том, что для тестируемого в проекте метода вопросы о том, насколько успешно удастся сопоставить глаголу в конкретном контексте некоторую модель управления, является принципиальным. Также результаты очень сильно зависят от множества выделяемых для данного глагола значений, от результатов разметки примеров в выбранной системе значений. Известно, что результаты ручной семантической разметки разными экспертами расходятся в 30% случаев.

Таким образом, решение задачи 4) потребовало целой серии экспериментов для разработки специальных технологий нахождения контекстных МУ. Исследование не предполагало использование полного синтаксического анализа. Однако эксперименты показали, что извлечение моделей управления непосредственно из примеров корпуса дает очень низкую точность за счет большого процента ошибок в распознавании самих моделей управления. Для того, чтобы ошибки такого рода не мешали основной задаче исследования, был введен дополнительный модуль квазисинтаксического анализа (подробнее см. в п.3.10).

Работа над задачей 5) показала, что в силу принципиальной несводимости значений, выделяемых различными словарными источниками, решение этой задачи требует ручной работы эксперта (более подробно см. п.3.9).

Третий этап представлял собой использование специализированного словаря Апресяна для обучения по размеченному корпусу. Он включал:

- 6) разработку методов доуточнения моделей управления и системы значений в результате обучения;
- 7) процедур, позволяющих получать МУ с семантическим расширением.

Четвертый этап исследования включал решение следующих задач:

- 8) применение метода снятия многозначности глаголов (уменьшение числа таксономических классов, не связанных с данной глагольной лексемой в конкретном контексте) к тестовому корпусу (подкорпусы НКРЯ со снятой и неснятой морфологической омонимией и с семантической разметкой).
- 9) оценка результатов экспериментов на 10 глаголах из исходной выборки.

Проведенные эксперименты можно разбить условно на 2 части. В первой части информация о моделях управления извлекалась из лексикографических словарей общего назначения (ОЖ, МАС). Возможности понижения степени многозначности оценивались относительно множества значений, выделяемых в соответствующих словарях. Поскольку эксперимент

дал очень низкие результаты и показал, что информация о моделях управления, извлеченная из словарей, требует уточнения на материале корпуса, то основные усилия были сосредоточены на второй части.

Во второй части источником информации о моделях управления послужил специализированный словарь глагольного управления (АПР). Возможность понижения степени многозначности оценивалась относительно набора семантических тэгов НКРЯ. Далее были проверены некоторые гипотезы относительно различных методов уточнения моделей управления на основе экспертной оценки данных словаря, а также на основе данных обучающего корпуса. Последняя из решаемых задач включала оценку возможностей снятия многозначности с учетом семантических характеристик существительных, входящих в МУ.

3.6 Основные принципы подготовки исходных данных. Ключевые глаголы и корпуса.

Для эксперимента было отобрано 400 глаголов из верхней части частотного списка (частотный список С.А. Шарова и данные корпуса со снятой омонимией). Надо отметить, что для ряда глаголов данные о частотном ранге по списку С.А. Шарова и по данным корпуса существенно расходились. Преимущественно выбирались глаголы, имеющие высокий ранг по частоте в обоих списках.

В выборку попали глаголы, имеющие не менее двух тэгов семантической разметки, используемой в НКРЯ. Мы опирались на данные НКРЯ, поскольку традиционные лексикографические источники часто выделяют слишком дробные значения. Часть глаголов, имеющих достаточно большое число значений в толковых словарях, не являются многозначными с точки зрения семантической разметки НКРЯ.

При отборе глаголов мы исходили из того, что глаголы одного семантического класса часто обладают похожей структурой актантов и имеют аналогичную структуру семантических переносов. Так, например, глаголы движения имеют один обязательный актант (в именительном падеже) и хотя бы один из локативных актантов (существительные, обозначающие направление движения, конечный пункт движения, исходный пункт движения и т.п.). Таким образом, отобранные глаголы должны были представлять разные семантические классы.

Для более подробного анализа и оценки точности были выбраны 10 высокочастотных глаголов из разных семантических классов (глаголы движения, каузации обладания, восприятия и т.п.). В это множество вошли глаголы с разным количеством и составом основных актантов и с разной степенью многозначности (от 3 до 8 тэгов из расширенного набора семантических тэгов НКРЯ)

Первоначально из корпуса со снятой омонимией были отобраны предложения с ключевыми

глаголами, за исключением контекстов, где данные глаголы стояли в форме причастий и деепричастий.

Тестовые корпуса, на которых происходила оценка, представляли собой случайные подвыборки из исходного корпуса со снятой омонимией для каждого глагола отдельно по 200 примеров (в отдельных случаях по 100 примеров) на каждый подкорпус. Для тестирования метода на корпусе с неснятой омонимией также случайным образом было отобрано по 200 примеров на один тестовый корпус. При выборе такого объема тестовых корпусов мы опирались на данные, приведенные в [15]. Авторы данной работы утверждают, что при распределении контекстов по частотности реализации значений достаточно разметки тестовой части контекстов, выбранных случайным образом, объем такой выборки может варьироваться от 100 до 300 контекстов.

3.7 Извлечение информации из толковых словарей. Особенности работы с неспециализированными лексикографическими источниками

При извлечении информации о модели управления из словарей учитывались следующие параметры:

- класс глагола по переходности;
- другие актанты – существительные и местоименные существительные в соответствующих падежах и соответствующими предлогами;
- уточнения, касающиеся других способов оформления актантов (инфинитивным оборотом, придаточным изъяснительным и т.п.);
- ограничения на форму ключевого глагола; общие семантические ограничения: одушевленность актанта (если она специальным образом оговаривалась в словаре), абстрактность vs. конкретность (как правило, данная информация в словаре отсутствовала). Учитывалась только информация о МУ, явным образом упомянутая в словарной статье.

Лексические ограничения учитывались при условии, что данные ограничения были указаны в словаре во фразеологической зоне.

В МАСе при глаголе выставляется специальная помета для переходности, а также внутри некоторых значений добавлена информация, указывающая на особенности управления для данного значения. Пример фрагмента словарной статьи см. рис. 1. (информация, извлекаемая в качестве модели управления) подчеркнута:

4. Нанести какой-л. удар, наградить (оплеухой, пощечиной, подзатыльником и т. п.). *Он вспыхнул и дал мне пощечину. Мы бросились к саблям.* Пушкин, Выстрел. ... || кому чем, по чему, во что или без доп. Прост. Ударить. *Дать по физиономии. Дать в ухо.*
□ *Контрорщик, которому прискучат шум, брань и причитывания, выскочит и даст кому-нибудь по уху.*

Рис. 1. Фрагмент словарной статьи для глагола *дать*

В толковом словаре Ожегова информация о глагольном управлении задается вопросительными словами, для отдельных значений указываются особые способы оформления актантов (например, придаточное предложение), морфологические ограничения на форму самого глагола (например, не 1 и 2 л.), ограничения на семантические признаки актанта. Семантические ограничения вычисляются по вопросительным словам: так актант, замещающий вопросительное местоимение *кто* получает семантические пометы anim&consp.

В результате каждому значению приписывалась соответствующая модель управления (см. таблица 1).

Таблица 1. Связь значения с МУ по словарям МАС и ОЖ. Фрагмент.

S&nom;S&acc	ОЖ1-1
S&nom;S&acc	ОЖ1-2
S&nom;S&acc	ОЖ1-4
S&nom;S&acc	ОЖ1-5
S&nom;S&acc	МАС 1-1
S&nom;S&acc	МАС 1-3

После извлечения всей информации из словарей формируются классы (кластеры) значений относительно некоторой МУ (см. таблица 2):

Таблица 2. Классы значений для глагола *найти*. Фрагмент.

МУ	Знач. Ож	Кол-во Ож	Знач. МАС	Кол-во МАС
S&nom	ОЖ2-4	1	1-3, 2-3, 2-5	3
S&nom; S&acc	ОЖ1-1, ОЖ1-2, ОЖ1-4, ОЖ1-5	4	1-1, 1-3	2

Множество значений, выделяемых в одном словаре, принципиально не сводимо к множеству значений в другом словаре. Как видно из таблицы 2, количество значений, связанных с некоторой МУ также не совпадает. Выработка единой системы значений требует ручной работы эксперта. В результате при оценке метода относительно конкретного словаря мы использовали систему значений соответствующего словаря.

3.8 Извлечение МУ из специализированного словаря

Третий из экспериментальных источников – словарь Апресян-Палл ([17]) – существенно более последователен в связывании МУ со значениями. Для каждого значения приводится достаточно полный список МУ. Каждая МУ представлена типовым примером. Однако словарь часто содержит очень дробные значения, что для задачи снятия семантической омонимии в корпусе нерелевантно. Для корпусной разметки важно учесть более крупные семантические классы.

С другой стороны, несмотря на дробность значений, под одним значением часто объединяются употребления глагола как с одушевленными, так и с

неодушевленными актантами в одном и том же значении, что расходится с общими принципами разметки значений в корпусе.

Большим преимуществом данного словаря является то, что некоторые актанты, например, обозначающие направление, местонахождение, источник и т.п., представлены обобщенными схемами. Такие актанты обычно выражаются предложно-падежными сочетаниями, при этом один и тот же актант может иметь более 3 способов выражения. По всему массиву словаря были извлечены типичные шаблоны для расшифровки обобщенных схем в моделях, в т.ч. списки стандартных предлогов для данного типа актанта с соответствующими падежами, списки типовых наречий, списки типовых сложных предложных сочетаний типа *справа от*.

3.9 Выделение "триад": базовая модель управления - значение по лексикографическим источникам - семантический класс по классификации Национального корпуса русского языка.

Данный этап был необходим для проведения серии экспериментов, в которых в качестве источников использовались словарь [Апресян-Палл] и результаты обучения по корпусу. В таксономической разметке корпуса семантические тэги привязаны к значениям по словарю Ожегова. Модели управления наиболее полно и последовательно описаны в словаре Апресяна. Попытки последовательно связать значения по Ожегову, корпусные тэги и данные словаря Апресяна лишний раз подтвердили тезис о том, что информация о значениях леммы из разных словарей не может быть автоматически сведена к единому списку значений.

В результате каждая двойка <МУ_j, s_i(Апр)> была связана со значением по словарю Ожегова и с корпусной пометой.

Связывание МУ с множеством значений позволяет сделать предварительную оценку «различительной» силы той или иной МУ. Так, некоторые МУ оказываются однозначно связаны только с одним значением. Например, МУ {S&nom; за+S&acc} для глагола *болеть* соответствует только одному значению глагола – переживать (класс эмоции). Другие МУ связаны практически со всеми значениями. Например, МУ {S&nom} для глагола *идти* связана со следующим набором корпусных тэгов {move, move_metaph, LF, exist}, а по данным МАС 16 из 20 значений связано с данной МУ. Средняя степень словарной многозначности для глаголов из выборки по словарю Апресяна составляет 5,5 значений на глагол, по тэгам НКРЯ 3,2 значения на глагол. Однако при такой оценке не учитывается сама частотность значений в реальном корпусе. Таким образом, уже на данном этапе видно, что возможность различить значения на основе МУ радикальным образом различается по глаголам. Для одних глаголов каждая из МУ связана не более чем

с двумя значениями. Для других глаголов МУ для всех значений одна.

3.10 Извлечение контекстных моделей управления для ключевых глаголов в корпусах

Задачей данного этапа являлось извлечь из каждого предложения соответствующего корпуса контекстную МУ, а также свести ее рядом преобразований (если это возможно) к одной из словарных моделей.

В результате серии экспериментов мы пришли к выводу о том, что таким оптимальным контекстом может служить отрезок предложения от знака препинания до знака препинания, включающий ключевой глагол. Более широкий контекст рассматривался только в особых случаях: когда одна из словарных моделей содержала актанта, выраженный придаточным предложением. В ряде ситуаций анализируемый отрезок был короче: в частности, при сочинении двух глаголов (ср. [*вытер руки*] и [*дал ему по затылку*]).

В первоначальном варианте МУ приписывались, исходя из следующего правила: предложение из корпуса проверяется на наличие полного набора элементов, входящих в соответствующую словарную МУ. Первыми проверялись более полные модели. То есть, модель с тремя актантами проверялась раньше, чем модель с двумя актантами, являющаяся подмножеством первой (например, {S&nom;S&acc;S&loc} раньше, чем {S&nom;S&acc}). Предложению приписывалась максимальная МУ. Данный метод дал неудовлетворительные результаты (около 40% точности) по следующим причинам:

- покрытие корпуса полными МУ составляет не более 60%, т.е. в 40% контекстов отсутствует один из обязательных актантов (например, существительное в именительном падеже);
- процент предложений, где отсутствующий актанта никак не выражен в пределах предложения, достаточно высок; в среднем в 8% контекстов на глагол содержался только сам глагол без актантов (предложения типа *Нашел. Дай.*).

Было принято решение разработать блок квазисинтаксических преобразований для уточнения неполных МУ.

В результате выделение контекстных МУ происходило с двух сторон. С одной стороны, из анализируемого отрезка извлекался «скелет», т.е. элементы, которые потенциально могли бы быть актантами глагола и некоторые служебные элементы (см. в п.3.1. определение МУ). С другой стороны, этот «скелет» подвергался преобразованиям, целью которых было, по возможности, свести его к одной из словарных МУ. Блок включал 3 типа правил: правила извлечения «скелета», правила извлечения и уточнения семантической информации, правила преобразования исходного контекста.

При извлечении «скелета» из анализируемого отрезка удалялись все словоформы, кроме относящихся к потенциальным актантам (существительных, предлогов и т.п.). Также удалялись нерелевантные для поиска актантов существительные и словосочетания с ними. Было выделено несколько типов «неактантных» элементов, включая устойчивые словосочетания типа *тем не менее, более того* и т.п., временные обстоятельства (например, *в это время*), некоторые предложные группы, которые могут выступать только как обстоятельства (например, конструкции со словами *ради, вопреки, по причине*). Такие элементы задавались настраиваемыми списками, которые могут легко редактироваться в процессе обучения. Отдельным блоком правил удалялись лишние «неактантные» существительные в родительном падеже.

В «скелете» сохранялась только часть информации, содержащейся в грамматической и семантической разметке. Для существительных это падеж, одушевленность и первый семантический тэг (abstr или concr), местоимения также получали семантические тэги в соответствии с их грамматическим разрядом и родом (например, местоимению *кто* в скелете соответствовал элемент S&nom&anim&concr, местоимению *оно* – S&nom&inan&no1, где no1 – обозначает concr или abstr). Отдельно обрабатывались квантитативные словосочетания (слово, обозначающее количество + сущ. в род.п.), а также предложные словосочетания типа *справа от* + сущ. в род.п., *в середине* + сущ. в род.п. В результате оставлялось только одно из двух существительных, которому приписывался падеж первого слова и семантические характеристики слова в род.п.

Полученный скелет подвергался дальнейшим преобразованиям. Учитывались регулярные изменения падежной формы актантов в ряде конструкций: например, дательный падеж для подлежащего актанта в инфинитивном обороте (*Мне хотелось побыстрее решить этот вопрос*), родительный падеж при отрицании (*Он не дал мне денег*) и т.п. В этом случае действовало правило замены данного падежа на каноническую форму (например, S&gen → S&acc).

Отдельный блок правил – правила интерпретации неполных контекстных МУ – отвечал за восстановление обязательных актантов, явным образом не выраженных в анализируемом отрезке (подробнее см. п. 3.11). Так, при отсутствии в отрезке существительного в именительном падеже оно восстанавливалось за исключением случаев, когда в список словарных МУ входила МУ без данного актанта. В результате применения последнего правила покрытие корпуса словарными моделями управления увеличилось до 83%.

Из всего множества выделенных в корпусных примерах актантов оставались только те, которые входили хотя бы в одну МУ из источника. При

использовании корпусного обучения оставлялись также некоторые другие актанты.

Часть ошибок применения правил (например, удаление S&gen) компенсировалась правилами уточнения неполных контекстных МУ. Часть ошибок в распознавании МУ приводила к увеличению среднего уровня многозначности, но не к потере точности.

3.11 Проблема неполных и «избыточных» контекстных МУ. Правила уточнения

После применения вышеописанных правил часть контекстных МУ (МУ, извлеченных из анализируемого предложения) не совпадали полностью ни с одной из словарных МУ. Уточнение таких МУ осуществлялось специальными правилами.

В случае, если несовпадающая МУ включала больше актантов, чем в любой словарной МУ, контексту приписывались все максимальные словарные МУ, которые входили в данную в качестве ее подмножества. Например, глаголу *идти* в предложении *Он шел от берега к дому* должно соответствовать 2 МУ: {S&nom;от+S&gen}, {S&nom;к+S&dat}, поскольку в словарях для данного глагола отсутствуют МУ из трех актантов.

Второй случай, когда контекстная МУ, не совпадала ни с одной словарной МУ, но являлась подмножеством одной или нескольких из них. Например, для глагола *дать* одна из контекстных МУ {по+S&dat} (ср., -- *А может, и по уху бы не дал.*) являлась подмножеством словарной МУ {S&nom; S&dat; по+S&dat}. Специальной группой правил такая МУ сводилась к одной из полных словарных МУ.

Более сложный случай, когда в контекстной МУ отсутствует прямое дополнение (S&acc), ср. *Вы нам не дадите на денёк?* (глагол имеет значение 'каузация обладания') и при этом существует словарная МУ, не содержащая прямое дополнение: ср. *Я ему как дам!* (значение физического воздействия). Здесь возможны следующие стратегии:

(1) контекстная МУ с отсутствующим актантом связывается только с соответствующим значением (со значением 'физическое воздействие' в примере);

(2) контекстной МУ приписывается объединение двух наборов значений (оба значения из примера).

В реальных текстах обязательный актант может быть опущен. Например, для глагола *дать* в значении 'каузировать обладание' прямое дополнение отсутствует в 17% контекстов. В результате первая стратегия может приводить к понижению точности. При втором подходе сохраняются более высокая степень многозначности.

Успешность одной или другой стратегии зависит от того, какова реальная структура значений и частотность МУ для конкретного глагола. То есть, решение, какое из правил уточнения контекстной МУ применять, зависит от свойств глагола, а также от необходимой точности и степени

многозначности. Выбор одного из правил также должен регулироваться на этапе обучения по корпусу (см. ниже).

3.12 Снижение степени многозначности с использованием информации, автоматически извлекаемой из источников

Эксперименты с неспециализированными источниками (МАС, ОЖ) показали, что ни один из источников не дает исчерпывающего списка МУ, покрытие корпуса МУ из каждого из словарей составляет в среднем 60% без применения правил восстановления отсутствующих актантов. При восстановлении отсутствующего именительного падежа покрытие составляет в среднем 80%, разброс от 70% до 100%. Покрытие 100% характерно для глаголов с базовой моделью управления {S&nom}. Например, для глагола *лежать* из словаря Ожегова извлекается только одна модель {S&nom}, при условии восстановления актанта в именительном падеже – покрытие 100%.

Разброс точности также достаточно велик: от 30% до 100%. Высокая точность достигается при очень низкой степени понижения многозначности. В частности, для глагола *лежать* с использованием словаря Ожегова многозначность не понижается, поскольку в этом словаре явным образом указывается только одна МУ для всех (12) значений.

Одним из принципиальных недостатков толковых словарей является то, что информация о необязательных актантах приводится, в основном, для периферийных значений. Из словаря невозможно извлечь исчерпывающего списка всех возможных оформлений факультативных актантов для некоторого значения. В ряде случаев это приводит к неудовлетворительным результатам. Так, например, для глаголов движения в первом значении не указываются возможные оформления актантов, обозначающих место (по словарю [Апресян-Палл] их более 20). Возможности разного предложно-падежного оформления учитываются только для переносных значений. Например, для глагола *идти* значение 'перемещаться' в МАС связано только с МУ S&nom, в то время как со значением 'соответствовать' (ср. *Платье ей к лицу*) соотносится МУ {S&nom;к+S&dat}. В корпусе данная МУ составляет 11% и соответствует первому значению в 60% случаев. В результате, если алгоритм работает по совпадению контекстной и словарной МУ, глаголу приписывается неправильное значение. Если же глаголу приписывать все значения, связываемые как с полной МУ, так и с МУ S&nom (только обязательный актант), степень многозначности понизится в незначительной степени.

В результате, в среднем понижение степени многозначности (WSR) составило 32%. Для подвыборки глаголов степень многозначности понизилась в более значительной степени, в среднем WSR составил 52% при точности 61%. Однако разброс данных по отдельным глаголам очень велик,

связан с особенностями глагольного управления, а также с непоследовательностью выделения значений и подачи информации о моделях управления в словарях. При этом достаточно большая доля ошибок приходится на контексты, где многозначность разрешается полностью.

Таким образом, можно сделать следующие выводы относительно возможности использования неспециализированных источников:

(1) информация о МУ, автоматически извлеченная из словарей и отражающая только грамматические характеристики актантов, дает очень низкие результаты;

(2) существенным недостатком словарей является то, что далеко не всегда можно извлечь полную базовую МУ непосредственно по грамматическим пометам; например, в МАС для глагола *дать* отсутствует информация о том, что при значении 'передать из рук в руки' в МУ должно/может входить существительное в дат.п. (при этом данный глагол относится к синтаксическому классу трехактантных глаголов);

(3) ни один из источников не дает исчерпывающей информации о множестве всех МУ, связанных с некоторым значением;

(4) модели управления, взятые из общих словарей как таковые, помогают снизить многозначность за счет фразеологизмов и менее частотных значений, связанных со специфической, не базовой МУ.

Данные анализируемых словарей можно использовать для относительно надежного отсека редких тэгов (аналогично отсекаемому разбору 'деепричастие' для предлога *для*).

Специализированный словарь [Апресян-Палл] дает значительно лучшие результаты (см. табл. 4.).

Таблица 4. Результаты использования специализированного словаря

	Покр.	WSR	WSD	точность
Апр	87%	53%	17%	86%

Задачей словаря являлось связать каждое значение со всеми возможными МУ для данного значения (ср. 25 МУ, включая расшифровку предложного оформления локативных актантов, для значения 'перемещение' глагола *идти*). Именно подробная информация о возможном оформлении необязательных актантов позволяет существенным образом понижать степень многозначности.

Несмотря на все преимущества такого специализированного словаря, существуют проблемы с его применением, аналогичные проблемам для МАС и словаря Ожегова: нет разделения актантов на обязательные и факультативные, не всегда последовательно выделяются значения, не все МУ и не все значения учтены. Таким образом, после автоматического извлечения МУ из словарей необходим еще этап уточнения МУ: либо экспертным путем, либо по обучающему корпусу.

Использование специализированного словаря

[Апресян-Палл] имеет ряд преимуществ по сравнению с использованием исключительно корпусных данных. Словарная информация может служить в качестве источника для определения основных (входящих в большое количество МУ) актантов и установления соответствий между кластерами значений и МУ. В случае, если одну синтаксическую позицию могут занимать актанты, имеющие разное оформление (как в случае локативных актантов), некоторые способы оформления могут быть низкочастотными в корпусе. Такие актанты не попадут в МУ при применении исключительно частотных критериев к выделению актантов. При этом, наличие их в МУ позволяет существенным образом понижать степень многозначности.

3.13 Использование обучающего корпуса

Привлечение данных обучающего корпуса улучшило характеристики метода:

средний процент уменьшения многозначности на 65%, точность 92% (при нежесткой системе обучения, т.е. если из множества значений, связанных с предложением, не удаляются наиболее редкие значения).

Уточнения касались следующих случаев:

- введение семантических тэгов, неучтенных в словаре АПР, выделение по размеченному корпусу идиоматических оборотов, уточнение некоторых лексических ограничений;
- уменьшение количества тэгов, приписанных глаголу в предложении, за счет отсека редких значений (не более 1% в обучающем корпусе);
- уменьшение количества тэгов для неполных извлеченных моделей ($МУ(s_i)$);
- уточнение множества МУ для некоторых значений.

Корпусные данные использовались для корректирования связей между МУ и значениями, а также для добавления неучтенных МУ. Изменения в связях между МУ и значениями происходило при следующих условиях:

- актант присутствовал в не менее, чем 10% примерах в обучающем корпусе и был связан с одним и тем же значением или группой значений;
- некоторое «правильное» значение было регулярно связано с данной МУ и при этом оно не входило в кластер, связанный с соответствующей МУ;
- некоторое значение из кластера, связанного с неполной МУ, ни разу не встречалось в числе правильных.

В первом случае МУ с таким актантом добавлялась в число словарных МУ. Во втором случае значение добавлялось в соответствующий кластер. В последнем случае соответствующее значение из кластера удалялось.

В результате оказалось, что существенный вклад обучающего корпуса касается уточнения неполных и избыточных контекстных МУ. Для многих

глаголов ситуация, когда по некоторым причинам в предложении не хватает актантов, оказывается более «благоприятной» для разрешения многозначности, чем полная стандартная модель. Так, например, множество возможных значений глагола *найти* в предложении с опущенными актантами уменьшилось с пяти максимально возможных до двух ('каузация обладания' и метафорический перенос по данному значению).

В силу неполноты исходной словарной информации после обучения объем некоторых кластеров значений, связанных с определенной МУ, увеличился. Однако это не повлияло на общее уменьшение степени многозначности и повышение точности.

3.14 Построение семантического расширения.

МУ и результаты применения расширенной МУ для понижения степени многозначности

Следующим диагностическим признаком, сужающим множество возможных значений, является семантический класс актанта. Добавление обобщенных семантических характеристик в МУ по обучающем корпусе позволило улучшить результаты еще на 7%-10%.

Точность оценки при этом понизилась на 3-5%. Это связано со следующими обстоятельствами:

- 1) с недостаточным объемом обучающего корпуса;
- 2) с ошибками в семантической разметке существительных;
- 3) с тем, что при расширении МУ семантическими характеристиками учитывались только первые значения существительных, при этом для некоторых классов существительных выбор характеристик *congr vs. abstr* зависит от семантики глагола, ср. например, существительное *журнал* в *порвать журнал vs. издавать журнал*.

Вклад семантической информации различается по глаголам. Так, для глаголов движения прямое значение физического перемещения характерно как для одушевленных, так и для неодушевленных объектов. Для глаголов восприятия наличие неодушевленного подлежащего невозможно при прямом значении. Данное правило можно отнести к разряду лексикографических эвристик, оно предсказуемо из общих представлений о том, как может быть устроена модель управления глаголов восприятия, даже если не использовать обучающий корпус и словарную информацию. В данном случае неодушевленный актант в именительном падеже автоматически имплицитно метафорический перенос. Однако для большинства эвристик находятся исключения, действие таких правил требуют дополнительной проверки.

Семантический признак актанта имеет неодинаковую различительную силу для разных типов актантов. Для обязательных актантов (S&nom и S&acc) признак *abstr vs. congr* часто разбивает кластер значений на прямое значение и множество переносных значений (метафоры и LF). Например, для глагола *вести* МУ {S&nom;S&acc} образует

кластер из 2-х значений. Для абстрактного актанта S&acc возможно только значение LF.

Семантические характеристики актантов не работают в случае, когда актанты выражены анафорическими местоимениями третьего лица *он, она, они*.

К сожалению, лексикографические источники не дают исчерпывающей информации о семантических ограничениях для каждого из выделяемых значений. Наш эксперимент показал, что данная информация, хотя и не дает полного разрешения многозначности (в некоторых случаях степень многозначности остается достаточно высокой: 2 значения на одно словоупотребление), тем не менее, позволяет существенно снижать многозначность, особенно для некоторых классов глаголов: глаголы обладания, восприятия, ментальные глаголы и т.п.

Информация о семантических ограничениях на актанты для каждого из выделяемых значений может быть получена:

- (1) с использованием некоторых лексикографических эвристик – фриквенталий;
- (2) извлечена из примеров, приведенных в словаре;
- (3) извлечена из обучающего корпуса;
- (4) с привлечением эксперта.

Способ (1) пригоден не для всех случаев, поскольку, во-первых, далеко не всегда семантические ограничения для конкретного глагола могут вытекать из общесемантических принципов. Во-вторых, мы не располагаем таким источником, в котором бы были достаточно четко и полно сформулированы такие эвристики. Классы исключений также требуют дополнительных исследований.

Третий способ имеет свои недостатки, поскольку примеров на некоторые редкие сочетания семантических характеристик разных актантов оказывается в небольшом обучающем корпусе очень мало, часто они периферийны или связаны с ошибочной семантической разметкой актантов. Некоторые сочетания могут быть не представлены в обучающем корпусе. В результате для редких комбинаций семантических признаков невозможно статистически достоверно разграничить случаи существенные, «диагностические» от случайных.

Была предложена следующая технология: первичная информация извлекается по примерам из словаря [Апресян-Палл]. Далее она корректируется с учетом данных корпуса. Вначале проверяются характеристики обязательных актантов (S&nom, S&acc). Если они не позволяют однозначно выделять значение, то семантические характеристики проверяются для каждой МУ отдельно. Эксперимент показал, что для более эффективного использования семантических характеристик корпусных данных не достаточно, при выбранном объеме корпуса менее частотные случаи требуют экспертного уточнения.

3.15 Выделение МУ и понижение степени многозначности в корпусе с неснятой морфологической омонимией

Использование данных корпуса с неснятой морфологической омонимией для тестирования исследуемого метода сталкивается с целым рядом трудностей. Во-первых, степень падежной омонимии для существительных достаточно высока. Дополнительный шум создается также из-за частеречной омонимии. Наибольшую проблему представляет омонимия союз *vs.* местоименное существительное, местоименное прилагательное *vs.* местоименное существительное, а также омонимия им. и вин. падежей для ряда глаголов, имеющих переходную и непереходную МУ. Если применение словарной информации для корпуса со снятой морфологической омонимией дает сравнительно невысокие результаты, то применение метода к корпусу с неснятой омонимией дает еще более низкие результаты относительно степени понижения многозначности.

Использовалась следующая технология приписывания МУ глаголу: из предложения извлекались все словоформы, имеющие хотя бы в одном из своих разборов тэг существительного (или местоименного существительного). Предложение связывалось с теми словарными МУ, которые входили в цепочку в качестве ее подмножества. Таким образом, предложение связывалось с несколькими МУ, что в ряде случаев увеличивало объем кластера значений, связанных с данным предложением. Для уточнения контекстных МУ также использовался квазисинтаксический модуль, включая связывание предлогов с зависимым существительным.

Необходимо отметить, что при этом результирующее понижение степени многозначности не так существенно отличается от результатов на корпусе со снятой морфологической омонимией, как ожидалось. Это связано с тем, что учет предложно-падежных групп позволяет понижать степень морфологической многозначности существительных. Квазисинтаксический блок также позволяет исключать некоторые комбинации падежных форм из числа возможных. Еще один положительный фактор – это то, что кластеры, связанные с разными МУ, являются вложенными множествами. В результате при связывании предложения с несколькими МУ количество приписываемых предложению значений не увеличивалось. В-третьих, при принятых правилах извлечения МУ большое количество альтернативных разборов для существительных игнорировалось, поскольку учитывалось только такое оформление актантов, которое было представлено в словаре.

При использовании словаря [Апресян-Палл] и результатов обучения WSR понизился в среднем на 10-15%.

3.16 Вклад информации о МУ глагола в понижение степени многозначности

Анализ результатов применения метода позволяет сделать целый ряд обобщений. Глаголы разбиваются на классы в зависимости от того, в какой степени грамматическая информация о МУ позволяет существенным образом понижать многозначность. Так для глагола *вести* МУ позволяет понизить степень многозначности с 5 до 1,8 значений, а с учетом семантических характеристик до 1,46 значения. С другой стороны, целый класс глаголов имеет одну МУ для всех значений (например, глагол *покинуть* имеет 1 МУ на 4 тэга).

Множество значений разбивается по словарным МУ на пересекающиеся кластеры. Количество элементов в кластере может колебаться от множества всех значений до одного (ср. МУ для *лечь* по ОЖ).

Актанты образуют иерархию с точки зрения их различительной силы. Их можно разбить на два класса:

(а) базовые актанты, такие как S&nom, S&acc, а также актанты, соответствующие семантическому классу глагола в его прямом значении (например, актант, указывающий на место, для глаголов движения; S&dat для глаголов типа *давать*);

(б) уточняющие актанты.

МУ, содержащая базовые актанты, приводится первой в словарных источниках. Базовые актанты наиболее частотны при данном глаголе в корпусе. Они, как правило, содержатся в нескольких МУ для данного глагола. Показательным является тот факт, что наибольшее количество значений имеют кластеры, связываемые с МУ, содержащей базовые актанты. Например, для глагола *дать* это будет модель <именительный, винительный, дательный>. Такая модель управления представлена почти во всех возможных для данного глагола значениях: прямое значение - класс 'каузировать обладание', лексические функции (*Ему дали работу*), класс 'физическое воздействие' (*Она дала ему пощечину*). При этом такая модель, как правило, имеет наибольшее покрытие (37% для глагола *дать*). Различить значения внутри такого кластера можно только с учетом семантической информации об актантах.

Второй класс актантов включает более специфические, необязательные актанты, например, предложные группы или инфинитив. Они обладают большей «различительной» силой. Наличие такого актанта в МУ существенным образом сужает множество соответствующих значений вплоть до одного, таким образом, он может служить диагностическим признаком для некоторого значения даже при отсутствии в контекстной МУ других актантов. Так, для глагола *болеть* актант за+S&gen в МУ задает только одно значение, относящее к классу 'эмоции'. Разное падежное оформление второго актанта при глаголах движения

также позволяет существенным образом сузить класс значений. Так, глагол *идти* имеет по разметке НКРЯ 8 тэгов. Для значения 'движение' возможно более 20 МУ. Однако каждая из этих МУ либо связана только с данным значением, либо максимальная величина кластера не превышает 3 значения. В кластер, связанный с МУ {S&nom}, входит 5 значений. Если в данном случае не включать вторые актанты в МУ, в один кластер (соответствующий МУ {S&nom}) попадут все значения глагола.

Основные ошибки данного метода связаны с тем, что в словарях не учитывается все множество возможных МУ, связанных с данной лексемой. Отдельную проблему представляют собой «ложные» актанты, т. е. случаи, синтаксической омонимии. Например, предложению *Он дал им по яблочку* будет поставлена в соответствие словарная МУ {S&nom;S&dat;по+S&dat} и соответственно глаголу в данном контексте будет приписано значение 'физическое воздействие'.

Таким образом, использование только грамматической информации о МУ для уточнения значения глагола имеет целый ряд ограничений. Одна и та же МУ может соответствовать достаточно большому количеству значений глагола. Для различения таких значений необходимо привлекать информацию о семантических свойствах актантов.

4. Выводы и обсуждение результатов

Итак, эксперимент показал, что ни один из выбранных лексикографических источников, включая и такой специальный источник как словарь моделей управления глаголов, не дает удовлетворительных результатов при автоматическом использовании извлеченной из них информации. Во-первых, как бы ни был полон список моделей управления для каждого из значений глагола, в реальном корпусе высок процент примеров, не покрываемых данными моделями. Значительная часть предложений содержит неполные МУ. Уточнение неполных МУ требует как введения дополнительного синтаксического модуля (или специального модуля приведения неполных МУ к словарным), так и обучения по корпусу.

Ни один из рассмотренных нами словарей не дает исчерпывающего списка моделей управления. Процент словарной полноты варьируется в зависимости от источника (т.е. от самого словаря) и от особенностей глагола. Словари также не содержат информацию о степени обязательности/факультативности актанта. Одно из основных препятствий использования словарной информации без предварительной обработки – это то, что в словарях не дается последовательно информация о всех возможных МУ, связанных с данным значением, что существенно снижает точность алгоритма. Возможно, объединение информации, извлекаемой из разных словарей,

позволит улучшить результаты.

Результаты применения данного метода существенным образом зависят от индивидуальных свойств глагола и его семантического класса. Таким образом, основные ограничения метода связаны с неполнотой словарной информации. Во-вторых, для целого ряда глаголов объем некоторых кластеров значений, связанных с определенной МУ достаточно велик, они могут включать большинство значений глагола. В-третьих, уменьшению числа значений мешает шум при выделении контекстной МУ (например, наличие обстоятельственных именных групп, с тем же оформлением, что и у возможных актантов).

В случае же, если мы хотим также учитывать семантические характеристики актантов, ни один словарь не дает исчерпывающей информации, даже если использовать в качестве данных материал словарных примеров. В данном случае уточнение возможно только с использованием обучающего корпуса.

Несмотря на указанные выше недостатки, использование информации о МУ для ряда глаголов позволяет уменьшить среднюю степень многозначности достаточно надежно. Наибольшее количество тэгов связывается с предложениями, в которых выделяются базовые модели управления (содержащие обязательные синтаксические актанты), поскольку стандартные метафорические переносы происходят с сохранением МУ, но с изменением семантического наполнения этой МУ.

Анализ данных показал, что признаки контекста образуют некоторую иерархию с точки зрения их различительной силы. Грамматические и семантические характеристики некоторых актантов позволяют однозначно приписывать значение глаголу независимо от свойств входящих в ту же контекстную модель других актантов. Наибольшей различительной силой обладают лексические ограничения (случаи, когда глагол приобретает данное значение только в устойчивом словосочетании), далее следуют периферийные (факультативные актанты), отсутствие одного из базовых актантов в МУ также служит диагностическим признаком. Одна из возможностей улучшения рассмотренного метода – выработка технологии, позволяющей построить такую иерархию по словарю и обучающему корпусу.

Мы предполагаем, что улучшение результатов может также дать использование примеров из словарей в качестве базового обучающего корпуса. Однако существенное понижение степени многозначности можно достичь только с учетом семантических характеристик актантов, а для ряда глаголов только с привлечением информации о более узком тезаурусном классе актантов.

5. Литература

- [1] Brown, P.F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert Mercer. Word-sense disambiguation

- using statistical methods. // ACL. – 1991. – V.29. – P. 264-270.
- [2] Dagan I., Itai A., Schwall U. Two languages are more informative than one // Proceedings of the ACL, 1991 (29). P. 130-137.
- [3] Fellbaum, Christian (ed.) WordNet: An Electronic Lexical Database. MIT Press, 1998.
- [4] Gale, William A., Church, Kenneth W. and Yarowski, David. A method for disambiguating word senses in a large corpus. // Computers and the Humanities. – 1992. - Vol. 26 – P. 415-439.
- [5] Gildea, Daniel, Daniel Jurafsky. [Automatic Labeling of Semantic Roles](#) // Computational Linguistics. – 2002. - Vol. 28. – No 3. – P. 245-288.
- [6] Johnson, C., Fillmore, C., Petruck, M., Baker, C., Ellsworth, M., Ruppenhofer, J., and Wood, E. FrameNet: Theory and Practice. 2002. [Electronic resource]. – 2002. – Mode of access: <http://www.icsi.berkeley.edu/framenet>.
- [7] Kingsbury, P., Palmer, M., and Marcus, M. Adding semantic annotation to the Penn TreeBank. // Proceedings of the Human Language Technology Conference HLT-2002 – San Diego, California. – 2002.
- [8] Lesk M. Automatic sense disambiguation using machinereadable dictionaries: How to tell a pine cone from a ice cream cone. // Proceedings of SIGDOC '86. New York. Association for Computing Machinery - 1986. – P. 24-26.
- [9] Lopatková, Markéta, Ondřej Bojar, Jiří Semecký, Václava Benešová, and Zdeněk Zabokrtský. Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. // Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors. Text, Speech and Dialogue: 8th International Conference, TSD 2005. – Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings, volume LNAI 3658. – Springer Verlag. – 2005. – P. 99-106.
- [10] Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. Chapter 7. – Cambridge, Massachusetts: The MIT Press. –1999. – P.230–262.
- [11] Mihalcea R., Chklovsky T., Kilgarriff A. Framework and results for English SENSEVAL // Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, July 2004, Barcelona. – Barseelona, Spain. – 2004. – P. 25–28. <ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-09.pdf>.
- [12] Ng H.T., Lee H.B. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach // Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96). – Santa Cruz. – 1996.
- [13] Scott Songlin Piao, Rayson P., Archer D., McEnery T. Comparing and combining a semantic tagger and a statistical tool for MWE extraction // Computer Speech & Language. – Vol. 19. – No 4. – 2005. – P. 378-397.
- [14] Shi, L., and Mihalcea, R. Semantic parsing using FrameNet and WordNet. // Proceedings of the Human Language Technology Conference (HLT/NAACL 2004). – Boston. – 2004.
- [15] Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Кобозева И.М., Нариньяни А.С., Селегей В.П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции Диалог'2004. – Москва – 2004. – С. 542–547.
- [16] Апресян Ю.Д. Лексическая семантика. – М.: «Наука», 1974. – 368 с.
- [17] Апресян Ю.Д., Палл Э. Русский глагол - венгерский глагол. Управление и сочетаемость. Будапешт, 1982.
- [18] 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. – М. – 2005. – С. 38-57.
- [19] Ляшевская О.Н., Кустова Г.И., Падучева Е.В., Рахилина Е.В.) Опыт семантического расширения морфологической разметки: таксономическая классификация лексики в национальном корпусе русского языка // НТИ, сер. 2. Информационные процессы и системы. – № 6. – 2005.
- [20] О.А. Митрофанова, В.В. Кадина, В.С. Савицкий. Экспериментальное исследование синтагматических свойств лексем на основе лексикографических описаний и корпусов текстов // Труды международной конференции MegaLing'2006– Горизонты прикладной лингвистики и лингвистических технологий. 20 - 27 сентября 2006 г., Украина, Крым, Партенит.
- [21] Рахилина Е.В., Ляшевская О.Н., Кобрицов Б.П., Кустова Г.И., Шеманаева О.Ю. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Лауфер Н.И., Нариньяни А.С., Селегей В.П. (ред.). Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». – 2006. – с. 445-450.

Extraction of verb argument structure from a dictionary in word sense reduction applications

Kobricov B.P., Lyashevskaya O.N., Toldova C.Ju.

This report deals with methods of word sense disambiguation (reduction) using the information about verb argument structure. Most of the systems based on this method require specially designed resources such as WordNet, FrameNet etc. We explore the possibility to extract and use the information available from the standard dictionaries including a Verb-argument dictionary. We used a subcorpus of National corpus of Russian language that has unambiguous morphological annotation as training and testing data. The aim was to reduce the number of tags for verbs in the semantic annotation. The experiment has shown that the information extracted from dictionaries could not be used as it is. All the tested dictionaries are inconsistent in the information about verb argument structure. However the extracted argument structure can be used as a seed set for future training. It allows to remove rare meanings and can reduce the number of semantic tags for a verb. The further corpus training and enriching the argument structure with general semantic properties of nouns can further improve the method.