УДК 004.85: 81.32

МЕТОД АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ МОДЕЛИ УПРАВЛЕНИЯ ГЛАГОЛОВ РУССКОГО ЯЗЫКА

Клышинский Э.С. *(klyshinsky@mail.ru)*,
Кочеткова Н.А. *(natalia_k_11@mail.ru)*,
Московский институт электроники и математики
Высшей школы экономики, Москва

Рассмотрен метод автоматической генерации словаря глагольного управления русского языка. Основой является статистическая обработка результатов поверхностного синтаксического анализа. Для анализа берутся неомонимичные группы слов, которые позволяют однозначно провести их синтаксический разбор. Для составления словаря использовался неразмеченный корпус объемом более 10 миллиардов словоупотреблений.

Введение

Глагольное управление является важной частью систем АОТ. Информация о глагольном управлении используется как для разрешения омонимии [Толдова 2008, Клышинский 2011], так и для снятия синтаксической неоднозначности [Гельбух 1999] и вообще синтаксического анализа [Волкова 2003]. В данной статье нас больше будут интересовать вопросы создания словаря глагольного управления.

Под моделью управления (МУ) глагола в данной статье будем понимать информацию о сочетаемостных характеристиках глагола. То есть, глагольное управление показывает какими грамматическими (род, число, падеж) или семантическими (связи слова в тезаурусе или онтологии) параметрами должно обладать существительное, для того чтобы быть связанным при помощи подчинительной связи с данным глаголом через заданный предлог или без него. В данной работе мы обрабатываем только лексическую информацию, то есть, словарь МУ будет содержать в себе информацию о том, в каком падеже может находиться существительное, присоединяемое к данному глаголу через заданный предлог. Заметим, что подобное определение не делает разницы между актантами и сирконстантами. Словарь сочетаемости показывает с какими словами может быть синтаксически связан данный глагол.

Существующие словари, составленные вручную, содержат в себе информацию высокого качества, но обладают малым объемом. Так, например, работа [Денисова 2002] содержит в себе всего 2500 статей, хотя и весьма представительных. Объем более ранних печатных работ [Розенталь 1986, Апресян 1982] также не очень велик. Электронный словарь «КроссЛексика» [Большаков 2011] содержит в себе около 2 млн. связей, но на его основе словарь моделей глагольного управления не составлялся. Более скромный словарь [Бирюк 2012] содержит 10 000 статей, но также недостаточен для программных решений. Часть приведенных выше словарей, вместе с информацией, доступной в НКРЯ, использовались для составления электронных словарей [Кобрицов 2007]. Развитием данного метода стал ресурс Фреймбанк [Framebank 2012], который хранит информацию не только о модели управления, но и о типе синтаксической связи – порядка 27 000 пар [Ляшевская 2009]. Таким образом, в проекте Фреймбанк хранится меньше информации, хотя она находится на качественно ином уровне.

Автоматическое извлечение словаря МУ уже предлагалось в ряде работ как для русского [Гельбух 1999], так и для других языков [Manning 1993, Messiant 2008, Preiss 2007]. В них предлагалось использовать конечные автоматы (КА) для распознавания отдельных цепочек «глагол + группа существительного» [Manning 1993] или системы синтаксического анализа [Гельбух 1999, Messiant 2008, Preiss 2007]. Далее из полученных результатов выделялись синтаксические связи между глаголом (а в случае [Preiss 2007], и существительным) и зависимыми словами. Недостатки синтаксического и морфологического анализа приводили к тому, что результат включал значительное число ошибок (от 5 до 20%), требующее ручного вмешательства.

Для устранения этой проблемы могут использоваться размеченные корпуса. На данный момент НКРЯ содержит 17,5 млн. предложений, содержащих почти 210 млн. словоупотреблений [НКРЯ 2012]. Омонимия снята лишь с 516 000 предложений, содержащих почти 6 млн. словоупотреблений. Самый крупный на данный момент синтаксически размеченный корпус СинТагРус по состоянию на 2011 год содержал более 45 000 синтаксически размеченных предложений в середине года [Frolova 2011] и более 49 000 предложений на конец года [СинТагРус 2011]. Заметим, что по нашим оценкам для составления словаря МУ для 25000-30000 глаголов требуется корпус примерно в 6 миллионов предложений, то есть объемов существующих корпусов всё еще недостаточно.

Метод генерации словаря

В отличие от предыдущих работ, предлагается использовать лишь неомонимичную часть текстов на русском языке. Омонимия в русском

языке имеет качественные отличия от омонимии, например, в английском. Для того чтобы убедиться в этом мы провели разметку корпуса текстов, результаты которой показаны в Таблице 1. Для сравнения были взяты две новостные дорожки за 2009 год: Компьюлента для русского языка и Рейтерс для английского, а также сайт Lenta.ru за 2005-2009 гг. Для разметки русской части использовалась система морфологического анализа «Кросслятор» [Елкин 2003, Школа 2011], для разметки английской части – библиотека Pymorphy. Также проводилась морфологическая разметка с использованием модуля АОТ, которая подтвердила полученные закономерности. То есть конкретные цифры отличаются для корпусов различной тематики и стилистики, используемых систем морфологического анализа, но приведенные цифры отражают разницу между омонимией в русском и английском языках.

Табл. 1.

Тип	**Lenta.ru**	**Компьюлента**	**Reuters**
Однозначные	46,28%	52,55%	38,87%
Неизвестные	4,38%	4,27%	7,65%
Неоднозначны по части речи	5,08%	14,88%	50,35%
Неоднозначны по параметрам	35,21%	24,68%	2,79%
Неоднозначны по норм. форме	4,67%	3,61%	0,32%

Для составления словаря МУ использовались КА для выделения синтаксически корректных в большинстве случаев групп. В качестве таких групп брались следующие. (1) Группа существительного, следующая за единственным глаголом в предложении, синтаксически подчиняется глаголу, прилагательные подчиняются существительному. (2) Единственная группа существительного в начале предложения перед единственным глаголом подчиняется данному глаголу, при этом прилагательные подчиняются существительному. (3) Прилагательные между предлогом и существительным подчиняются существительному. (4) Если после существительного, находящегося не в родительном падеже, следует существительное в родительном падеже, причем между первым существительным и глаголом находится предлог, то второе существительное подчиняется первому. (5) В конструкции «глагол + предлог + существительное/глагол» третье слово можно считать существительным и оно подчиняется глаголу. (6) Группа вида «предлог + прилагательное + прилагательное/существительное + существительное», в которой согласование возможно только когда второе слово считается прилагательным, считается группой существительного и все прилагательные подчиняются последнему слову.

Составление словаря МУ ведется по следующему алгоритму.

Шаг 1 – извлечение сочетаний из текста. На данном шаге проводится морфологическая разметка корпуса текстов. Далее по шаблонам 1 и 2 отбираются глагольные сочетания. В формировании данной базы принимают участие существительные, однозначные по части речи, но, возможно, неоднозначные по падежу. В связи с этим полученные сочетания в чистом виде еще не могут использоваться для создания словаря глагольных МУ. Проиллюстрируем предложенный метод на примере. Пусть из текста извлечены, среди прочего, следующие связанные сочетания вида «глагол (+предлог) +существительное».

состоятся вечера
приглашает на концерт
исполнят произведения
состоится встреча
примут участие
откроется выставка

Шаг 2 – составление базы сочетаемости слов. Полученные на предыдущем шаге сочетания приводятся к нормальной форме, после чего рассчитывается их встречаемость. Из полученной базы отсеиваются сочетания, встретившиеся меньше заданного количества раз. В примере показаны абсолютные значения встречаемости указанных сочетаний, приведенных к нормальной форме.

ПРИГЛАШАТЬ;НА;КОНФЕРЕНЦИЯ;218
ПРИГЛАШАТЬ;НА;КОНЦЕРТ;281
ПРИГЛАШАТЬ;НА;КОНЬЯК;3
ПРИГЛАШАТЬ;НА;КОРАБЛЬ;17
ПРИГЛАШАТЬ;НА;КОРДОН;3
ПРИГЛАШАТЬ;НА;КОРОНАЦИЯ;6

Шаг 3 – составление модели управления предлогов. Из сочетаний, полученных на Шаге 1, отбираются те, в которых существительное однозначно по падежу. Для них рассчитывается встречаемость пар вида «предлог+существительное в заданном падеже». Для удобства все пары с одним предлогом собираются в единую запись. Всего была получена информация о примерно 300 предлогах (в том числе и составных). Для нашего примера получим записи следующего вида:

*К 0*0*8950*21*17*5*
*НА 0*0*0*30707*0*89*
*ПЕРЕД;0*0*0*0*5*0*

Цифры показывают, сколько раз данный предлог встретился с существительными в им., род., дат., вин., твор. и предл. падежах соответственно. Цифры здесь взяты для глагола «приглашать», но на практике суммировались значения для разных глаголов.

Из модели управления предлогов вручную был отсеян шум, связанный с ошибками. Анализировалась частотная информация, полученная для предлогов, а не вся модель управления в целом. Так предлог «К» никогда не сочетается с винительным, творительным или предложном падежами. Таким образом, мы получили информацию о том, в каком падеже может встречаться существительное, связанное с данным предлогом.

*К 0*0*8950*0*0*0*

*НА 0*0*0*30707*0*89*

*ПЕРЕД;0*0*0*0*5*0*

Также отсеивались сочетания предлога с именительным падежом. Как показала практика, несмотря на то, что подобное сочетание возможно (например, «идти в солдаты» или сочетания с «а-ля»), но в получаемой базе в подавляющем большинстве случаев является шумом.

Шаг 4 – получение модели управления для глаголов. Из базы, полученной на шаге 1, строим базу сочетаний вида «глагол + предлог + падеж существительного», после чего отбрасываем все варианты, запрещенные моделью предлога (шаг 3) или встречающиеся только один раз (шаг 2). Искомый словарь содержит записи следующего вида.

ПРИГЛАШАТЬ;
*17063*2103*863*14439*1583*41*
*ПРИГЛАШАТЬ;БЕЗ;0*97*0*0*0*0*
*ПРИГЛАШАТЬ;В;0*89*0*17847*0*1775*
*ПРИГЛАШАТЬ;ВМЕСТО;0*9*0*0*0*0*
*ПРИГЛАШАТЬ;ВО;0*0*0*1005*0*29*
*ПРИГЛАШАТЬ;ВОЗЛЕ;0*5*0*0*0*0*
*ПРИГЛАШАТЬ;ВОПРЕКИ;0*0*9*0*0*0*
*ПРИГЛАШАТЬ;ДЛЯ;0*1129*0*0*0*0*
*ПРИГЛАШАТЬ;ДО;0*1*0*0*0*0*
*ПРИГЛАШАТЬ;ЗА;0*0*0*817*25*0*
*ПРИГЛАШАТЬ;ИЗ;0*297*0*0*0*0*
*ПРИГЛАШАТЬ;ИЗ-ЗА;0*61*0*0*0*0*
*ПРИГЛАШАТЬ;ИЗО;0*5*0*0*0*0*
*ПРИГЛАШАТЬ;К;0*0*8950*21*17*5*
*ПРИГЛАШАТЬ;КО;0*0*489*0*0*0*
*ПРИГЛАШАТЬ;КРОМЕ;0*9*0*0*0*0*
*ПРИГЛАШАТЬ;НА;0*0*0*30707*0*89*
*ПРИГЛАШАТЬ;НАД;0*0*0*0*21*0*
*ПРИГЛАШАТЬ;ОТ;0*17*0*0*0*0*
*ПРИГЛАШАТЬ;ПЕРЕД;0*0*0*0*5*0*
*ПРИГЛАШАТЬ;ПО;0*0*509*0*0*65*
*ПРИГЛАШАТЬ;ПОД;0*0*0*117*19*0*
*ПРИГЛАШАТЬ;ПОДОБНО;0*0*5*0*0*0*
*ПРИГЛАШАТЬ;ПОСЛЕ;0*189*0*0*0*0*
*ПРИГЛАШАТЬ;ПОСРЕДИ;0*9*0*0*0*0*
*ПРИГЛАШАТЬ;ПРИ;0*0*0*0*0*1*
*ПРИГЛАШАТЬ;РАДИ;0*17*0*0*0*0*
*ПРИГЛАШАТЬ;С;0*0*1*0*815*0*
*ПРИГЛАШАТЬ;СО;0*1*0*0*18*0*
*ПРИГЛАШАТЬ;СОГЛАСНО;0*0*13*0*0*0*
*ПРИГЛАШАТЬ;ЧЕРЕЗ;0*0*0*109*0*0*

Фильтрация результатов для присоединения существительного без предлога не проводилась так как для автоматической фильтрации отношения цифр недостаточно. Для некоторых глаголов стандартные ошибки оказывались более частотными, чем корректные варианты. Анализ полного списка глаголов вручную при этом не проводился.

Шаг 5 – фильтрация базы сочетаемости. Из базы, полученной на шаге 1, отсеиваем все сочетания, не подходящие под полученную на шаге 4 модель управления. Одновременно может быть устранена падежная неоднозначность. В итоге мы получаем базу сочетаемости глаголов с существительными.

Результаты экспериментов

Мы взяли корпус объемом 10,5 млрд. словоупотреблений: подкорпус беллетристики (Библиотека Мошкова – 688 млн словоупотреблений, lib.rus.ec – 8,9 млрд), новостные тексты за 1999 – 2011 гг. (РИА Новости – 220 млн, Коммерсант и Независимая газета – по 99 млн, Взгляд – 95 млн и др., всего около 800 млн), научных текстов (авторефераты, диссертации, статьи, всего более 60 млн). Было выделено более 23 млн. уникальных сочетаний «глагол+предлог+существительное». Из полученной базы применением первых трех шагов алгоритма было извлечено около 425 000 сочетаний «глагол + предлог + разрешенные падежи». Качество полученных результатов находится на уровне не ниже 95% для всех сочетаний «глагол + предлог + падеж» и около 99% для однословных предлогов (слова, входящие в составной предлог, могут использоваться и в качестве именной группы). Оценка пропусков в модели не проводилась.

Заметим, что в полученный словарь глагольного управления не попала значительная часть слов. Исследование результатов показало, что в выделенных связках «глагол+предлог+существительное» приняло участие 22200 глаголов из 26400, представленных в использованном морфологическом словаре, и 55200 из 83000 существительных. Слова не были включены в словарь по одной из двух причин. Во-первых, это редко встречающиеся слова, имеющиеся в морфологическом словаре. Во-вторых, это существительные, омонимичные прилагательным во всех своих формах, например, «белый», «красный», «больной». Заметим, что введение шаблонов 5 и 6 позволило сократить количество подобных слов.

Кроме того, полученные сочетания совершенно не различают лексических значений глаголов. Несмотря на это, наличие подобного словаря сочетаний позволяет перейти к практическому решению целого ряда задач в анализе текстов: снятие омонимии, фильтрация результатов синтаксического анализа и др. Кроме того, наличие большой базы позволит попытаться перейти и к задачам, связанным с семантикой.

Сравнение новых результатов с предыдущими [Клышинский 2011], полученными на корпусе в 7,5 млрд словоупотреблений, показывает, что происходит насыщение словаря. При увеличении корпуса на 30% число глагольных сочетаний выросло лишь на 15%. Сравнение вклада разных корпусов показало, что беллетристика из lib.rus.ec и новостная лента вдвое меньшего объема дают сопоставимый вклад: 5,4 млн сочетаний «существительное + прилагательное» против 4,1 млн и 21,6 млн комбинаций «глагол + существительное» против 14,3 млн. Следовательно для извлечения МУ из текстов более важен стиль изложения и богатство лексики, чем объем корпуса.

Для сравнения результатов был использован словарь [Бирюк 2012] из которого были выбраны глаголы, сочетающиеся со словом «пример» без

предлога. В связи с большим количеством сочетаний, полученных в базе, анализ проводился только для глаголов на буквы А, Б и В. В словаре [Бирюк 2012] было найдено только два сочетания: «брать пример» и «взять пример». В нашем словаре было найдено 139 сочетаний, из них 104 встретились более одного раза, а 77 – больше двух. Десять самых частотных сочетаний связаны со следующими глаголами: быть (встретилось 18706 раз), видеть (1204 раз), воспользоваться (1043 раза), брать (785 раз), вспомнить (726 раз), бывать (532 раза), взять (468 раз), вдохновлять (418 раз), встречаться (403 раза), вдохновляться (192 раза). Данный пример показывает, что полнота полученного словаря сочетаний значительно выше. Из проанализированных сочетаний сомнения вызвало только одно «ВЕЛЕТЬ ПРИМЕР» (по всей видимости «велеть примером показать что-то»). Помимо этого были обнаружены сочетания глагола с «на пример» (искаженное «например»), количество которых не превысило 1% от всех сочетаний со словом «пример». Подобное сочетание является синтаксически корректным, хотя и грамматически неверным.

В заключении заметим, что описанный метод был применен для текстов на английском языке [Гурбанов 2012] и показал хорошие результаты. При применении методов снятия омонимии он показывает сопоставимую полноту выделения сочетаний, хотя это несколько понижает качество результатов.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 10-01-00800). Авторы выражают признательность сотрудникам корпорации «Галактика» Антонову А.В. и Баглею С.Г. за проведенные вычислительные эксперименты.

Список литературы

[Апресян 1982] Апресян Ю.Д., Палл Э. Русский глагол - венгерский глагол. Управление и сочетаемость. Будапешт, 1982.

[Бирюк 2012] Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. Словарь глагольной сочетаемости непредметных имен русского языка // электронная публикация, режим доступа – http://dict.ruslang.ru/abstr_noun.php

[Большаков 2011] Большаков И.А., Гельбух А.Ф. Большой электронный словарь как политематический справочник и формирователь запросов к Интернету // Материалы международной конференции «Диалог 2011», 2011 г. сс. 124-134

[Волкова 2003] Волкова И.А., Мальковский М.Г., Одинцев Н.В. Адаптивный синтаксический анализатор // Труды конференции «Диалог 2003» сс. 401–406.

[Гельбух 1999] Гельбух А. Разрешение синтаксической неоднозначности и извлечение словаря моделей управления из корпуса текстов // Материалы VIII Международной конференции KDS-99 (Крым - 13-18.09.1999г. - Кацивели)

[Гурбанов 2012] Гурбанов Т.П., Клышинский Э.С. Параллельный алгоритм составления словаря глагольного управления для новостных текстов на

английском языке // Сб. трудов 15 научно-практического семинара «Новые информационные технологии», М., 2012.
[Денисова 2002] Словарь сочетаемости слов русского языка / Под ред. П.Н. Денисова, В.В. Морковкина. 3-е изд., испр. М., АСТ, 2002. 816 с.
[Елкин 2003] Елкин С.В., Клышинский Э.С., Стекляников С.Е. Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003, т. 1, Дивноморское. 2003
[Клышинский 2011] Клышинский Э.С., Кочеткова Н.А., Литвинов М.И., Максимов В.Ю. Метод разрешения частеречной омонимии на основе применения корпуса синтаксической сочетаемости слов в русском языке // Научно-техническая информация. Сер. 2: Информационные системы и процессы. № 1 2011 г., сс. 31-35
[Кобрицов 2007] Кобрицов Б.П., Ляшевская О.Н., Толдова С.Ю. Снятие семантической многозначности глаголов с использованием моделей управления, извлеченных из электронных толковых словарей // Интернет-математика – 2007, электронная публикация, режим доступа – http://download.yandex.ru/IMAT2007/kobricov.pdf
[Ляшевская 2009] Ляшевская О.Н., Кузнецова Ю.Л. Русский фреймнет: к задаче создания корпусного словаря конструкций // Труды конференции «Диалог 2009» сс. 306-312.
[НКРЯ 2012] Статистика. Национальный корпус русского языка // Электронная публикация, режим доступа – http://ruscorpora.ru/corpora-stat.html
[Розенталь 1986] Розенталь Д.Э. Управление в русском языке // М.: Книга, 1986 г.
[СинТагРус 2011] Итоговый отчет о работе по программе Фундаментальных исследований Президиума РАН «Корпусная лингвистика» 2011 г. «Разработка системы синтаксического анализа русских текстов на базе корпуса «СинТагРус»» // электронная публикация, режим доступа – http://corpling-ran.ru/files/I/boguslavsky.pdf
[Толдова 2008] Толдова С.Ю., Кустова Г.И., Ляшевская О.Н. Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы // Труды конференции «Диалог 2008», С. 522–529.
[Школа 2011] Материалы Летней студенческой школы компьютерной лингвистики // электронный ресурс, режим доступа – http://clschool.miem.edu.ru/Материалы-школы.html
[Framebank 2012] Поисковый интерфейс системы Fraemebank // электронный ресурс, режим доступа – http://framebank.ru/
[Frolova 2011] Frolova T.I., Podlesskaya O.Yu. Tagging lexical functions in Russian texts of SynTagRus // Труды конференции «Диалог 2011», сс. 207-218
[Manning 1993] Automatic Acquisition of a Large Subcategorization Dictionary from Corpora // In Proc. of the 31st Meeting of ACL, pp. 235–242.
[Messiant 2008] Messiant C., Korhonen F., Poibeau T. LexSchem: A Large Subcategorization Lexicon for French Verbs // In Proc. of LREC 2008
[Preiss 2007] Preiss J., Briscoe T., Korhonen A. A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora // in Proc. of the 45 Annual Meeting of the Association of Computational Linguistics, pages 912-919.