Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов

Automatic construction of word combination database using a huge text corpus

Клышинский Э. С. (klyshinsky@mail.ru),
Кочеткова Н. А. (natalia_k_11@mail.ru),
Литвинов М. И. (promithias@yandex.ru)
Московский государственный институт электроники и математики

Максимов В. Ю. (vadimmax2000@mail.ru)
Институт прикладной математики им. М. В. Келдыша РАН

В статье рассмотрены вопросы автоматического формирования базы сочетаемости слов (глагол или деепричастие + существительное, прилагательное + существительное, причастие + существительное) на основе анализа размеченного корпуса большого размера — более 109 словоупотреблений. Данная работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы.

Введение

Информация о взаимном сочетании слов является достаточно важной для задач анализа текстов на естественном языке. Обладая подобной информацией, можно, например, существенно повысить качество и скорость синтаксического анализа, причем как глубинного, так и поверхностного. Подобная база может использоваться для определения меры близости текстов, их содержания, при снятии омонимии и так далее.

Наиболее часто для этих целей используются коллокации, то есть устойчивые словосочетания. Устойчивость подобных сочетаний во многом является трудно определяемой величиной. В различных задачах под ними могут пониматься фразеологизмы, идиомы и несвободные словосочетания [1]. В связи с этим для практических задач используются статистические методы определения связности соседних слов, например, MI [2], t-score [3] и некоторые другие. Полученные таким образом коллокации могут использоваться для составления информационных портретов [4], снятия омонимии [5], выделения понятий предметной области [6], кластеризации документов, генерации связанных текстов и целом ряде других задачах.

Однако важную роль при анализе текстов играют и свободные словосочетания. Подобная информация также может помочь при решении многих задач. В связи с этим отечественные лингвисты уже длительное время ведут работы по созданию подобных словарей. На данный момент разработаны весьма представительные словари, как в бумажном [7], так и в электронном виде [8]. Однако объем подобных словарей с точки зрения машинной обработки текста прискорбно мал. Так, например, [7] содержит в себе всего 2500 статей, хотя и весьма представительных, приводящих не только информацию о сочетании слова с другими, но и толкования данного слова, его грамматические характеристики. В работе [8] приводится более 10 000 статей, что охватывает порядка 3–5 % современной русской морфологии. В области составления словарей коллокаций также объемы полученных результатов не слишком велики. Так, например, в работе [5] указывается, что был составлен словарь на 30 000 коллокаций. Зачастую это связано с тем, что составляются словари конкретных предметных областей и задач, а работы для языка в целом практически не проводятся.

Временные затраты на создание подобных словарей достаточно велики, что собственно и объясняет небольшой объем. В связи с этим встает во-

прос автоматизации процесса создания подобных словарей. Кроме того, для их применения в машинной лингвистике необходимо приведение словарей к представлению, удобному для машинной обработки. При этом вопрос стоит о привлечении и изучении больших объемов текста.

Ранее уже предпринимались попытки извлечения информации о глагольном управлении и глагольном примыкании из больших корпусов, а также для других видов сочетаемости слов. Однако работа с большим корпусом путем его просмотра требует огромного количества времени. Так, например, работы, проводимые Большаковым И. А. в течение 20 лет, позволили ему получить базу сочетаемости для 185 тысяч слов и выражений, в том числе около 57 тыс. титулов словаря для существительных (раздельно для единственного и множественного числе) и 38 тыс. титулов для глаголов (раздельно для инфинитива и личных форм). Общее количество сочетаний превышает 1,75 млн [9].

Объем и сложность проведенных работ объясняется среди прочего наличием различного рода неоднозначностей. Их наличие требует проведения ручной или автоматизированной (но ни в коем случае не автоматической) разметки, по результатам которой из текста выделяются, например, глаголы и зависимые от них слова. Дополнительные трудности вводит синтаксическая неоднозначность, в связи с чем для каждого предложения необходимо предварительно построить дерево зависимостей (хотя бы и мысленно), и на его основании пополнить базу глагольного управления.

Метод выделения сочетаемости слов

Для практических задач зачастую хватает информации о том, что данный глагол может употребляться с данным существительным. Для автоматического определения таких связей необходимо решить проблему лексической и синтаксической неоднозначностей. Для этих целей были выдвинуты две гипотезы. Первая из них (как нам представляется — наиболее сильная) состоит в том, что в тексте достаточно большого объема группы из однозначных с точки зрения морфологического анализа слов будут встречаться достаточно часто, чтобы собрать статистически значимые результаты. Под однозначностью здесь мы понимаем случай, когда в результате морфологического анализа слова возвращается единственная строка его нормальной формы. В связи с тем, что в русском языке для большинства слов имеется достаточное количество форм, вероятность обнаружить однозначное слово относительно велика. A priori основной вопрос заключался в том, насколько часто в тексте будут встречаться группы подобных слов. Вторая гипотеза состояла в том, что некоторые группы слов могут быть синтаксически однозначно подчинены другим словам даже без проведения синтаксического анализа. В соответствии со сформулированными гипотезами для генерации базы глагольной сочетаемости нами были использованы следующие простые положения.

1. Следующая за единственным глаголом группа существительного синтаксически подчиняется данному глаголу.
2. Единственная группа существительного, расположенная в начале предложения перед единственным глаголом, синтаксически подчиняется данному глаголу.
3. Прилагательные, расположенные перед первым в предложении существительным или между глаголом и существительным, синтаксически подчиняются данному существительному.
4. Положения 1–3 могут быть применены к деепричастиям и причастиям.
5. В тексте на русском языке должно быть представлено достаточно большое количество неомонимичных групп.

Само расположение выделенных групп с большой (но не стопроцентной) вероятностью позволяет говорить о корректности определения синтаксических зависимостей. Отсутствие неоднозначности гарантирует корректность определения нормальной формы слов. И, наконец, корпус текстов большого объема может гарантировать статистическую значимость результатов.

Итак, для рассмотрения были отобраны синтаксические конструкции, включающие глагол и единственную группу существительного перед ним или первую группу существительного после него. При этом группа существительного описывалась следующим образом: предлог притяжательное_местоимение числительное группа_прилагательных существительное. Все части группы существительного являются необязательными, а притяжательные местоимения и числительные игнорируются. Также отбрасывались и наречия. Точность результатов при этом будет определяться точностью выбора неомонимичных слов, корректностью выбора последовательности слов и вероятностью правильного применения второй гипотезы. Представительность результатов определяется объемом анализируемого корпуса и вероятностью встретить неомонимичную группу с заданными характеристиками.

Итак, для создания базы сочетаемости слов необходимо проанализировать корпус текстов большого размера, выделяя из него последовательности слов, отвечающие предложенным шаблонам. Для каждой уникальной последовательности должна быть подсчитана ее встречаемость, которая в дальнейшем используется для определения статистической значимости результата.

Описание эксперимента

Для экспериментов использовался полученный ранее и несколько обновленный и расширенный корпус текстов [10]. В качестве основы корпуса были использована Библиотека Мошкова, включающая в себя порядка 680 млн словоупотреблений. Кроме того, была использована еще одна коллекция художественной прозы, объемом около 120 млн словоупотреблений, включающая в себя как классических, так и современных авторов. Также использовалась новостная лента, опубликованная на сайтах РБК, Лента.ру, Российская и Независимая газеты, РИА Новости (всего более 325 млн. словоупотреблений), и новостные ленты околокомпьютерной тематики Компьюлента.ру и PCWeek (37 млн словоупотреблений). Конкретный объем каждого из источников приведен ниже в таблице. Общий объем корпусов составил почти 1,2 млрд словоупотреблений.

Источник	Объем, млн словоупотреблений
Библиотека Мошкова	680
РИА Новости	156
Доп. корпус прозы	120
Независимая газета	89
Лента.ру	33
Российская газета	29
PCWeek	28
РБК	21
Компьюлента	9
Итого	**1165**

Все полученные комбинации слов сохранялись в базе данных, работа с которой заняла основное время эксперимента. Для морфологического анализа использовался модуль морфологического анализа «Кросслятор» [11].

По результатам экспериментов были получены базы сочетаемости глаголов и существительных, деепричастий и существительных, существительных и прилагательных, существительных и причастий. Для каждого указанного типа сочетаний подсчитывалось общее количество их вхождений, то есть сколько раз в корпусе встретились сочетания данного типа. Кроме того, подсчитывалось количество уникальных сочетаний данного типа. Объем получившихся баз приведен ниже в таблице. Числитель показывает общее количество обнаруженных вхождений, знаменатель — количество уникальных сочетаний. Дополнительно подобный подсчет был осуществлен для сочетаний, встретившихся в корпусе более одного и двух раз (третий и четвертый столбец, соответственно).

Исследование результатов показало, что в выделенных парах приняло участие 21 500 глаголов из 26 400, представленных в морфологическом словаре, 53 300 существительных из 83 000, представленных в морфологическом словаре и 23 700 прилагательных из 45 300 имеющихся. Большое количество глаголов объясняется гораздо меньшей степенью их омонимичности. Низкое количество прилагательных объясняется тем, что из нескольких прилагательных, стоящих перед существительным, в базу помещалось только первое.

Пара	Всего вхождений, млн	> 1 повторения, млн	> 2 повторений, млн
Глагол+сущ.	65 / 8,3	60,3 / 3,5	57,7 / 2,3
Деепр.+сущ.	3,5 / 0,88	2,8 / 0,31	2,6 / 0,18
Сущ.+прил.	9,9 / 1,3	9,2 / 0,56	8,8 / 0,36

Наибольшая повторяемость сочетаний была достигнута на новостных текстах. Наиболее часто встречающими сочетания глагола и существительного оказались следующие (слова приведены к нормальной форме, рассматривается встречаемость во всех формах).

Сочетание	Встречаемость
Сообщить РИА	624 691
Передавать корреспондент	327 903
Покачать голова	304 597
Принять участие	271 250
Иметь в вид	201 167
Принять решение	140 090
Говориться в сообщении	132 385
Сообщать агенство	118 615
Сказать собеседник	115 959
Идти речь	108 306

Наиболее часто встречающимися глаголами в различных сочетаниях стали следующие (результаты округлены до тысяч).

Сочетание	Встречаемость
Быть	2 590 000
Сказать	1 908 000
Сообщить	1452 000
Иметь	721 000
Применять	623 000
Получить	525 000
Заявить	515 000
Передавать	462 000
Идти	450 000
Сообщать	425 000

Среди сочетаний существительное + прилагательное наиболее часто встречающимися оказались следующие.

Сочетание	Встречаемость
Ближайший время	23 664
Правый рука	19 809

Сочетание	Встречаемость
Официальный представитель	19489
Последний время	18555
Военный служба	17933
Большой количество	17737
Официальный сайт	17385
Левый рука	16121
Информационный агенство	15503
Молодой человек	14699

Наиболее часто встречающимися существительными в парах прилагательное + существительное в различных сочетаниях стали следующие.

Сочетание	Встречаемость
Время	69438
Голос	61409
Человек	58313
Рука	54467
Жизнь	52125
Система	49855
Количество	48821
Свет	46438
Место	42973
Работа	41741

Полученные результаты до некоторой степени соотносятся с имеющимися данными о частотах распределения слов русского языка. В значительной мере здесь чувствуется влияние новостной лексики.

Анализ показал, что в результаты не попали принципиально неоднозначные слова, такие, например, как «красный», выступающий как в роли прилагательного, так и в роли существительного. Кроме того, в базу не вошли устаревшие и чрезвычайно редко употребимые слова, например, «взгревать», «издаиваться», «парагвайка» и так далее.

Следует заметить, что в связи с неоднозначностью в рамках данного эксперимента в базу не попали целые пласты сочетаний. Так, например, омонимичными является большое количество предлогов, например, «при» (повелительное наклонение единственного числа от «переть»), «для» (деепричастие от «длить») и так далее. Однако подобная ситуация может быть исправлена достаточно легко введением фильтров. Автоматически брать все слова, которые могут быть предлогами, как предлоги было бы не всегда корректно. Так, например, слово «сверху» может выступать как в роли предлога, так и наречия, причем примерно равновероятно. С другой стороны, языковые конструкции, в которых данное слово встречается, существенно отличаются в зависимости от того, применяется в них наречие или предлог. В связи с этим в конечный результат будут включаться только «правильные» конструкции. Резюмируя можно сказать, что данный вопрос нуждается в дальнейшем исследовании.

Выборочный просмотр результатов показал, что количество ошибок не превышает 1 %. В области наиболее частотных сочетаний ошибки метода составляют порядка 0,1 %, тогда как сочетания, встретившиеся только один раз, выделяются с примерно 1–2 % ошибок. Часть из ошибок объясняется не совсем корректной обработкой некоторых видов конструкций. Так, например, в предложении «Хочу от лица коллектива поздравить юбиляра» конструкция «от лица» ошибочно относилась к глаголу «хотеть». Отдельную проблему представляют ассоциации, гиперболы и другие выразительные средства литературного языка. Так, например, конструкция «секретарша ускакала» хотя и выделяется правильно, но с трудом может быть названа характерной для поведения секретарей. С другой стороны, подобная конструкция встретилась в корпусе всего один раз и с очевидностью находится ниже уровня статистической значимости. Будучи оторванными от контекста, подобные конструкции удивляют, хотя их выделение с точки зрения приведенных выше шаблонов проводится вполне корректно. С другой стороны, в текстах одной из новостных лент регулярно встречались выражения вида «сообщает в четверг». Несмотря на серьезные возражения о стилистической корректности подобного сочетания, его выделение проводилось вполне корректно. Будучи же приведенным к нормальной форме («сообщать в четверг»), выражение не вызывает никаких возражений.

Наиболее частотные сочетания, полученные подобным образом, хорошо коррелируют с результатами, получаемыми методами выделения коллокаций (здесь автор хотел бы выразить признательность Ягуновой Е. В. и Пивоваровой Л. М. за проведенное сравнение). Так, например, в обоих методах самым встречающимся сочетанием в новостных текстах было «агентство сообщать».

Выводы

Приведенный в работе метод позволяет на больших объемах текстов получить приемлемые результаты по извлечению глагольной сочетаемости. Несмотря на то, что для построения баз было использовано около 1,5 % всех словоупотреблений, большой объем корпуса позволил получить представительный результат.

Проведенные эксперименты показали, что выдвинутые гипотезы вполне корректны, хотя и носят вероятностный характер. При этом точность получаемых результатов составляет порядка 99 %. Отдельной темой для исследований является корректность встречающихся конструкций с точки зрения правил, принятых в языке.

Полученный корпус глагольной сочетаемости позволит перейти к следующим экспериментам в области сочетаемости слов: снятие неоднозначностей, группировка слов по семантическим признакам и так далее. Планируется провести аналогичные эксперименты на корпусе со снятой статистическими методами омонимией. Это позволит значительно увеличить базу обрабатываемых сочетаний и, как следствие, существенно увеличить объемы базы. С другой стороны, это должно привести к увеличению процента ошибок.

Литература

1. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Сб. статей «Инструментарий русистики: корпусные подходы». — Хельсинки, 2008. С. 343–357
2. *Church K., Hanks, P.* Word association norms, mutual information, and lexicography, *Computational Linguistics*, 1990, 16(1), P. 22–29.
3. *Stubbs, M.* Collocations and semantic profiles: On the cause of the trouble with quantitative studies, 1995. *Functions of Language*, 1.
4. *Антонов А. В., Ягунова Е. В*. Лингвистический анализ информационного портрета как свертки множества текстов. Постановка эксперимента // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах-13». М.: МИЭМ, 2010. С. 50–59.
5. *Невзорова О. А., Невзоров В. Н., Зинькина Ю. В., Пяткин Н. В.* Интегральная технология разрешения омонимии в системе анализа текстовых документов «ЛоТА» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» М.: Изд-во РГГУ, 2007, С. 422–427
6. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э. Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» М.: Изд-во РГГУ, 2007, С. 70–75
7. *Словарь* сочетаемости слов русского языка / Под ред. П. Н. Денисова, В. В. Морковкина. 3-е изд., испр. М., АСТ, 2002. 816 с.
8. *Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю.* Словарь глагольной сочетаемости непредметных имен русского языка — http://dict.ruslang.ru/abstr_noun.php
9. *Большаков И. А.* Кросслексика — большой электронный словарь сочетаний и смысловых связей русских слов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15). — М.: РГГУ, 2009. 620 с.
10. *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15).– М.: РГГУ, 2009, С. 165–169.
11. *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