К вопросу о репрезентации данных о сочетаемости в электронных лексикографических ресурсах

М.В. Хохлова[1], А.М. Попов[2]

[1] Санкт-Петербургский государственный университет, [2] ООО Инфо-Кьюбс

m.khokhlova@spbu.ru, hedgeonline@gmail.com

Аннотация

В статье дается обзор существующих систем, представляющих информацию о сочетаемости. К ним относятся разнообразные словари, а также специализированные базы данных и другие ресурсы. Также затрагиваются вопросы, связанные с реализацией проекта по созданию интегрированной базы данных, которая содержит автоматически извлеченные коллокации и дополнительную информацию. В системе будут представлены примеры, полученные как на основе корпусов при помощи автоматических методов, так и на материале словарей русского языка. Ресурс может быть использован в разнообразных задачах прикладной лингвистики, в том числе связанных с автоматической обработкой текстов.

Ключевые слова: сочетаемость, коллокации, словари, база данных, корпусы текстов, статистика

1. Введение

Сведения о сочетаемости лексических единиц традиционно представлены в словарях и справочниках. При этом с появлением больших корпусов текстов и методов их обработки открылись новые возможности для репрезентации данных подобного рода. Прежде всего имеются в виду количественные подходы, которые хотя и стали использоваться применительно к русскоязычному материалу довольно давно (см., например, первые исследования [1]), однако получили новый импульс в связи с развитием информационных технологий. Таким образом, появилась возможность описать сочетаемость с точки зрения ее двусторонней природы, имеющей как языковой, так и вероятностный характер.

2. Обзор проектов

2.1. Словарные источники

Информация о лексической и синтаксической сочетаемости различных слов обычно описана в словарях (толковых, специализированных и др.), реже в языковых грамматиках. Для русского языка можно назвать целый ряд толковых словарей, в которых сочетаемость отражена довольно подробно [2-4 и др.]. Тем не менее, можно отметить, что не существует единой концепции описания данных в разных источниках. В толковых словарях информация, указывающая на ограниченную сочетаемость, представлена несколькими способами: знак ромба обозначает устойчивые словосочетания и фразеологизмы в словаре [3] (в нем приводятся данные о 13 тыс. подобных сочетаний при общем объеме словника более 80 тыс. единиц), в то время как в [4] для первых используется он же, а вторые вводятся при помощи знака тильды. Сочетаемость может не выделяться специальным образом, а приводиться в цитатах или речениях. Сами словарные статьи также имеют

разную структуру. Так, для слова «надежда» словари [3, 5] перечисляют только следующие устойчивые словосочетания «возлагать надежду», «питать надежду», «подает надежду» и «льстит себя надеждой». Однако эти примеры не включают другие словосочетания, которые тоже являются свойственными этой лексеме (например, «оправдывать надежды» или «вселять надежду»), поэтому специализированные словари могут служить источником дополнительной информации. Для русского языка такие словари существуют и являются великолепными справочными ресурсами, представляющими сочетаемость [6, 7]. Необходимо отметить, что словарь [4] является в некоторой степени уникальным, так как словарные статьи содержат особый справочный отдел, в котором приводятся сведения из других словарных источников. В основном они касаются первой фиксации слова или изменений, произошедших в его произношении или написании. Также существует уникальный лексикографический проект под руководством Ю.Д. Апресяна «Активный словарь русского языка» [8], который включает обширную информацию о сочетаемости, которая выделяется отдельно в словарных статьях. Материал отлично структурирован и включает сведения о синтаксических актантах, коллокациях и конструкциях. Тем не менее, словари по-разному отражают сочетаемость и покрывают примеры, поэтому так важно рассмотрение отличных друг от друга источников.

2.2. Электронные ресурсы и базы данных

Методы, применяемые во многих задачах прикладной лингвистики, можно разделить на две большие категории: использующие правила и реализующие статистические алгоритмы. Применительно к задаче автоматического определения сочетаемости можно сказать, что системы, основанные на первом подходе, появились раньше. В некоторой степени они напоминают словари, представленные в электронном виде. Информация о сочетаемости может быть получена в них для определенных моделей и с отсылкой к корпусу. Второй подход реализован в значительно меньшем количестве систем и подразумевает автоматическое извлечение сочетаемостной информации статистическими методами из данных большого объема. Далее выделенные словосочетания могут сопровождаться количественными характеристиками, позволяющими оценить степень устойчивости (или воспроизводимости) конструкций.

В основном работы, посвященные описанию сочетаемости и ее последующему представлению в специализированных базах данных, на протяжении долгого времени затрагивали англоязычный материал. Так, был разработан словарь «The Pattern Dictionary of English Verbs», который базируется на методологии Corpus Pattern Analysis, предложенной П. Хэнксом [9] и включает семантико-синтаксические шаблоны глагольного управления с иллюстрациями (словосочетаниями и предложениями). Уникальным проектом является ресурс FrameNet, созданный Ч. Филлмором [10]. В нем представлена информация о валентности и о семантических ролях для английского языка Общее число примеров, эксплицирующих употребление лексических единиц, превышает 200 тыс. предложений. Проекты, выполненные в русле этого подхода, существуют для испанского, китайского, корейского, немецкого, французского, шведского и японского языков, а также для бразильского варианта португальского языка. Также можно назвать проект “Collocations” Ланкастерского университета, который посвящен определению сочетаемости у пар синонимов на материале корпуса BNC [11]. Эта же функция реализована в ряде корпусов текстов английского языка, разработанных М. Дэвисом и доступных на его портале. Пользователь имеет возможность искать слова, находящиеся в одном и том же контекстном окружении с ключевым словом (функция “compare”), так и проверять сочетаемость у близких по значению единиц [12].

Для английского языка существует лексическая база данных DANTE [13], в которой описаны свыше 40 тыс. наиболее частотных слов. Каждое значение лексической единицы проиллюстрировано автоматически собранными цитатами из двухмиллиардного корпуса.

Статистический механизм лежит в основе подхода к представлению сочетаемости в системе Sketch Engine [14, 15]. В ней можно получить информацию о контекстном окружении лексем для многих языков в виде таблиц, моделирующих сочетаемость и соответствующих заранее определенных синтаксическим моделям. Автоматически выделенные словосочетания сопровождаются количественными оценками, указывающими на силу синтагматической связи. Также существует многоязычный проект SkELL (Sketch Engine for Language Learning) [16], в котором представлены заранее отобранные цитаты, репрезентирующие словоупотребление. В рамках данного проекта представлены корпусы для учебных целей, в которых представлены «чистые» тексты и наиболее удачные примеры.

Что касается других языков, то отдельно можно выделить систему DWDS (Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart) для немецкого языка, которая объединяет словарь и базу данных [17]. Система IDS, разработанная в Институте немецкого языка, предоставляет возможность просматривать более чем 220 тыс. коллокационных профилей [18]. Выдаваемые коллокации сопровождаются статистической оценкой при помощи коэффициента логарифмического правдоподобия. Также для каждого примера приводится наиболее характерная синтаксическая модель словосочетания. Для словенского языка была разработана база данных коллокаций [19]. В ней представлены более 44 тыс. словосочетаний и 150 тыс. примеров.

На сегодняшний день также существует ряд проектов, ориентированных на материал русского языка и направленных на статистическое исследование лексической и синтаксической сочетаемости на основе корпусов текстов. К ним относятся, например, база данных «Лексикограф» [20], база сочетаемости FrameBank, в которой представлены лексические конструкции [21], и словари, созданные на основе НКРЯ [22]. Среди последних можно назвать словарь глагольной сочетаемости непредметных имен русского языка и словарь русской идиоматики. Первый проект основывается на понятии лексических функций, введенном в работе [23], и описывает более 10 тыс. словосочетаний следующих моделей: 1) существительное + глагол; 2) глагол + существительное; 3) глагол + прилагательное + существительное. Есть возможность проводить поиск по значению, синтаксическому отношению, фазовому значению и оценке. Тем не менее, в словаре отсутствует количественная характеристика силы данной сочетаемости, которая была бы весьма актуальна для исследователей.

Проект CoSyCo ориентирован на создание базы данных синтаксических конструкций, в которой на данный момент представлены именные и глагольные словосочетания [24]. Поиск единиц осуществляется в корпусах разных функциональных стилей: художественном, научном и публицистическом. Каждое опорное слово (по которому производится поиск) снабжается списком синтагматических партнеров с их частотами.

3. База данных

3.1 Вводные замечания

На основании анализа имеющихся систем, можно сказать, что существует необходимость в интегрированном ресурсе, который бы давал доступ к информации о сочетаемости, полученной при помощи разнообразных квантитативных методов на материале корпусов текстов, а также сопровождал бы словосочетания ссылками на традиционные лексикографические ресурсы. Таким образом, речь идет о совмещении двух упомянутых ранее подходов.

После проведенного анализа толковых и специализированных словарей русского языка был сделан вывод о том, что необходимо также ввести единый формат представления информации о сочетаемости, так как словарные статьи в разных источниках имеют разную

структуру. Например, в едином формате могут быть указаны части речи входящих в словосочетание единиц, а также информация о том, в каких словарях они встретились.

3.2. Структура базы данных

На данном этапе работы была произведена морфологическая разметка текстов, включающая неоднозначность (см. пример разбора для словосочетания «по словам министра»). Так, единица «министра» анализируется как существительное в родительном или винительном падежах. Также мы решили рассматривать многокомпонентные единицы как единое целое (например, «по словам»). Это даст возможность пользователю в качестве выходных данных рассматривать одну единицу, не разделенную на слова, что важно для целого ряда задач.

<token start="7608" end="7610" value="по" lemma="ПО" tag="W,Prep"/>
<token start="7608" end="7610" value="по" lemma="по" tag="Prefix"/>
<token start="7611" end="7617" value="словам" lemma="СЛОВО" tag="W,Noun,dat,neu,pl,in,NP"/>
<token start="7608" end="7617" value="по словам" lemma="по словам" tag="Prnt"/>
<token start="7618" end="7626" value="министра" lemma="МИНИСТР" tag="W,Noun,gen,mas,sg,an,NP"/>
<token start="7618" end="7626" value="министра" lemma="МИНИСТР" tag="W,Noun,acc,mas,sg,an,NP"/>

В ходе работы над проектом коллокации нами извлекались в два этапа. На первой стадии нами были разработаны правила, которые описывали русскоязычные словосочетания и были применены к извлечению данных. Следующие синтаксические модели представлены в базе данных: 1) ADJ+N; 2) N+N; 3) V+N; 4) V+Prep+N. Второй этап включал такие статистические метрики как MI, t-score, log-likelihood и другие меры ассоциации [25]. Статистические данные использовались для оценки извлеченных словосочетаний.

Как уже указывалось, лексикографические ресурсы предоставляют полезную информацию о сочетаемости. Поэтому при создании системы мы планируем использовать данные из ряда словарей [2-6]. Будет использована информация, иллюстрирующая контексты слов (например, речения, цитаты, указания на ограниченную сочетаемость). Такие данные также позволят оценить и верифицировать коллокации в базе данных.

В качестве материала при разработке базы данных были привлечены разнообразные тексты. Были использованы новостная коллекция, специальные тексты (техника) и беллетристика. Также были выполнены эксперименты на доступных русскоязычных данных следующих корпусов:: ruTenTen, Aranea Russicum Maximum и НКРЯ.

Для нашего исследования нами была разработана специализированная база данных MySQL для хранения пар слов и их корреляционных значений согласно разным коллокационным мерам, содержащая три основных таблицы: таблицу слов; таблицу коллокаций; таблицу метрик.

Таблица слов состоит из троек, каждая тройка занимает свой ряд и хранит следующую информацию: уникальное слово (в нашем случае словарная форма, т.е. лемма), тег части речи и их совместная частота (сколько раз данная пара «слово-часть речи» встретилась в рассматриваемом корпусе).

Таблица коллокаций содержит частотную информацию о каждой паре ряда из таблицы слов, включая информацию о линейном порядке (это означает, что сочетание двух слов с противоположным порядком будет рассмотрено как два разных сочетания и независимыми частотами).

Ниже показана схема базы данных (см. рисунок 1).

Также имеется доступ ко всей необходимой количественной информации (размер корпуса, частота биграммы и независимые частоты двух отдельных слов), которая необходима для вычисления мер связанности, хранящихся в третьей таблице. Она

содержит значения статистических мер для каждой коллокации (ряда в таблице коллокаций), словосочетание может иметь иного значений метрик, вычисленных заранее и хранящихся в базе данных для исследовательских целей. Каждая метрика снабжена уникальным значимым строковым ID (обычно название метрики).

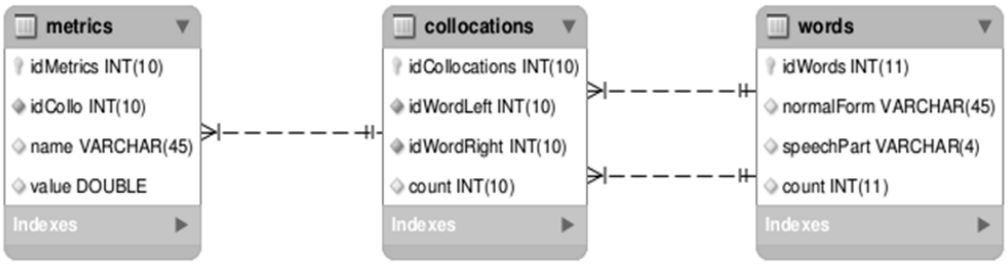

Рис. 1. Пример схемы базы данных

Простой SQL запрос предоставляет средства для извлечения необходимой информации с требуемыми ограничениями. Например, пользователь может ограничиться определенными частями речи и/или типами метрик.

Таблицы слов и коллокаций заполняются заранее созданной нами программой, записывая выходной результат обработки корпуса прямо в базу. Таблица мер, с другой стороны, может быть заполнена позже и даже увеличена, когда требуется ввести новую коллокационную метрику в базу. Также возможно вычислить некоторые метрики онлайн при помощи формулы, инкорпорированной в SQL запрос select.

4. Заключение

Следующим шагом могут быть методы машинного обучения применительно к задаче автоматического определения сочетаемости на материале языковых данных.

В статье мы попытались проследить основные принципы, лежащие в основе базы данных. Главная проблема заключается в необходимости интегрированного ресурса, который включит данные из словарей и корпусов текстов, снабженные достаточным количеством примеров для каждого слова. База данных использует MySQL и уже доступна пользователям на сайте в сети Интернет. Она содержит три таблицы, каждая из которых хранит информацию о словах, коллокациях и статистических мерах соответственно. Ресурс включит примеры из ряда корпусов (как специализированных, так и общих), статистическую оценку выделенных коллокаций при помощи мер ассоциации, ссылки на другие ресурсы, такие как словари и корпусы текстов. Разрабатываемая база данных может быть использована при обучении русскому как иностранному, для создания приложений, связанных с автоматической обработкой текстов, и при составлении словарей. В качестве будущей перспективы мы планируем добавить оценку пользователей для выделенных коллокаций.

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-2513.2018.6 «Исследование методов автоматического извлечения лексических конструкций на основе машинного обучения».

Литература

[1] Марков А.А. Об одном применении статистического метода // Известия Императорской Академии Наук. Серия VI. 1916. Т. 10, № 4.

[2] Словарь современного русского литературного языка: В 17 т. / Под ред. В.И. Чернышёва. М., Л.: Изд-во АН СССР, 1948—1965.
[3] Словарь русского языка: В 4-х т. / АН СССР, Ин-т рус. яз.; Под ред. А.П. Евгеньевой. 2-е изд., испр. и доп. М.: Русский язык, 1981—1984.
[4] Большой академический словарь русского языка: В 30 т. / Под ред. К.С. Горбачевича. СПб: Изд-во «Наука», 2004.
[5] Большой толковый словарь русского языка: А-Я / РАН. Ин-т лингв. исслед.; Сост., гл. ред. канд. филол. наук С.А. Кузнецов. СПб: Норинт, 1998.
[6] Борисова Е.Г. Слово в тексте: Словарь коллокаций (устойчивых сочетаний) рус. яз. с англо-рус. слов. ключевых слов. М.: Филология, 1995.
[7] Словарь сочетаемости слов русского языка / Под ред. П.Н. Денисова, В.В. Морковкина. – 3-е изд., испр. М., 2002.
[8] Активный словарь русского языка. Т. 1. А—Б / Отв. ред. акад. Ю.Д. Апресян. М.: Языки славянской культуры, 2014.
[9] Hanks P. Mapping meaning onto use: a Pattern Dictionary of English Verbs. AACL, Utah 2008.
[10] FrameNet. URL: https://framenet.icsi.berkeley.edu/fndrupal (дата обращения: 25.05.2018).
[11] Getting Collocations. URL: https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l08_3.htm (дата обращения: 25.05.2018).
[12] Corpora created by Mark Davies. URL: https://corpus.byu.edu (дата обращения: 25.05.2018).
[13] Database of Analyzed Texts of English. URL: http://www.webdante.com/index.html (дата обращения: 25.05.2018).
[14] Sketch Engine. URL: http://www.sketchengine.co.uk (дата обращения: 25.05.2018).
[15] Kilgarriff A., Baisa V., Bušta J,, Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. The Sketch Engine: ten years on. Lexicography. 1. 2014. P. 7-36.
[16] SkELL – examples and collocations for learners of English. URL: https://www.sketchengine.eu/skell/ (дата обращения: 25.05.2018).
[17] Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. URL: https://www.dwds.de/ (дата обращения: 25.05.2018).
[18] Kookkurrenzdatenbank CCDB. URL: http://corpora.ids-mannheim.de/ccdb/ (дата обращения: 25.05.2018).
[19] Slovene Lexical Database. URL: http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza (дата обращения: 25.05.2018).
[20] Проект «Лексикограф». URL: http://lexicograph.ruslang.ru (дата обращения: 25.05.2018).
[21] FrameBank. URL: http://framebank.ru/ (дата обращения: 25.05.2018).
[22] Словари, созданные на основе Национального корпуса русского языка. URL: http://dict.ruslang.ru/ (дата обращения: 25.05.2018).
[23] Бирюк О.Л., Гусев В.Ю., Калинина Е.Ю. Словарь глагольной сочетаемости непредметных имен русского языка. http://dict.ruslang.ru (дата обращения: 25.05.2018).
[24] Kopotev M., Escoter L., Kormacheva D., Pierce M., Pivovarova L., Yangarber R. CoCoCo: Online Extraction of Russian Multiword Expressions // Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing (10–11 September 2015, Hissar, Bulgaria). Sofia: INCOMA Ltd, 2015. P. 43–45.
[25] Manning Ch., Schütze H. Foundations of Statistical Natural Language Processing. Massachusetts: MIT Press, 1999.

On the Representation of Collocability in Online Lexicographic Resources

M. Khokhlova[1], A. Popov[2]

[1] Saint-Petersburg State University, [2] Info-Qubes

The paper gives an overview of the existing systems that represent collocability. Among these one can name various dictionaries and specialized databases and other resources. The authors also pay attention to the issues related to the project on building an integrated database that includes automatically extracted collocations and additional information. The system will comprise both automatically extracted examples and ones from Russian dictionaries. The present tool can be used in a wide range of tasks of applied linguistics, e.g. natural language processing.

Keywords: collocability, collocations, dictionaries, database, text corpora, statistics