АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81'322.2

Ю. С. Акинина, И. О. Кузнецов, С. Ю. Толдова

Сравнение двух методов автоматического извлечения участников события из неструктурированных источников*

Описывается одна из задач извлечения информации из текста, а именно – извлечение фактов (событий разного типа) из неструктурированных источников. Рассматриваются различные методы определения существительных, которые являются типовыми наименованиями участников события. Для решения данной задачи предлагается использовать статистические методы выделения коллокаций. Исследуются два подхода: выделение коллокаций на основе простой контекстной близости существительных к глаголу и выделение коллокаций на основе синтаксических связей существительного с глаголом. В результате сравнения двух методик авторы приходят к выводу, что вопреки первоначальной гипотезе о том, что информация о синтаксических связях должна обеспечить более точное и полное выделение участников и других характеристик событий, первая методика дает о них более полное представление. Анализ результатов показывает, что для успешного применения синтаксического фильтра необходимо учитывать опосредованные синтаксические связи.

Ключевые слова: *коллокации, глагольная сочетаемость, автоматический синтаксический анализ, корпусные методы*

ВВЕДЕНИЕ

Проблема извлечения информации из текста для автоматического/ автоматизированного пополнения онтологии

В последнее время в сфере информационных технологий все более острой становится проблема работы с большим объемом разнородных и слабоструктурированных данных. На передний план выходит задача упорядочивания, систематизации накопленных в различных областях деятельности знаний. В этой связи акцент в обработке контента смещается с простых задач информационного поиска на задачу предоставления пользователю некоторой обобщенной структурированной информации, полученной на основе агрегации информации из первичных источников. Для решения этой задачи были разработаны принципы и технологии семантического Веба. В основе идеологии семантического Веба лежит идея записи информации в виде семантической сети с помощью онтологий. Данная задача особенно актуальна для развития разных областей науки и техники, поскольку представление о важнейших достижениях и инновациях в некоторой научной сфере является неотъемлемым условием получения научных результатов в этой области.

В рамках развития направления семантического Веба во всем мире ведется активная разработка онтологий различных научно-технических областей. С одной стороны, такие онтологии разрабатываются при участии экспертов в данных областях. С другой стороны, поскольку в сети появляются все новые текстовые сообщения о каких-то событиях, достижениях, мероприятиях в соответствующих отраслях науки и техники, встает задача автоматизированного или автоматического пополнения таких онтологий информацией, извлеченной из неструктурированных источников. Таким образом, задача извлечения информации о некотором событии из текста оказывается востребованной для автоматического пополнения онтологий.

* Работа проводилась при финансовой поддержке Министерства образования и науки Российской Федерации в рамках государственного контракта № 07.524.11.4005 от «20» октября 2011 г., заключенного между Министерством образования и науки Российской Федерации и ЗАО «Эвентос».

Если мы имеем дело с неструктурированным текстом, то достаточно часто основным носителем информации о типе события является глагол (1) либо устойчивое словосочетание 'Глагол+Существительное' (2), а находящиеся рядом с ними существительные нередко обозначают типовых участников события (1):

(1) *Новый* ***материал*** *из углерода* ***синтезировали*** ***ученые*** *из университета Тунцзи в Шанхае.*

(2) *Потенциально нанотрубки* ***имеют*** *невероятно широкую* ***сферу применения.***

В примерах выделение используется для обозначения названий событий, а выделение и подчеркивание – для обозначения типовых участников или существительных, входящих в устойчивый оборот. А существительное *Тунцзи*, например, не является ни типичным актантом или обстоятельством места для события из (1), ни второй частью устойчивого словосочетания, как в (2).

Таким образом, выделение именно "типических" пар 'V–N' (глагол-существительное), во-первых, поможет автоматизировать построение онтологии экспертом, поскольку позволит выявить типовых участников того или иного события (а также их роли), во-вторых, оказывается полезным в задаче тезаурусного расширения соответствующих узлов онтологии: выделения синонимов, гипонимов и т.п.

Задача выделения 'V–N' коллокаций

Представляется, что задача нахождения именно типичных, а не случайных участников события может быть решена с использованием технологий выделения коллокаций. Задачу выделения коллокаций можно условно разделить на два этапа:

1) отбор кандидатов в коллокаты;

2) ранжирование пар по степени связанности.

Существует много методов как для отбора кандидатов, так и для их ранжирования. В настоящей работе используется один из стандартных методов ранжирования (PMI). Вопрос о критериях выбора метода остается за рамками настоящего исследования. В центре внимания находится вопрос о выборе кандидатов в коллокаты.

С одной стороны, существительные, обозначающие типичных участников, должны устойчиво встречаться достаточно близко к соответствующему глаголу. То есть можно предположить, что для извлечения интересующих нас 'V–N'-коллокационных пар при сравнительно большом объеме корпуса достаточно простой информации о совместной встречаемости существительного с глаголом в пределах некоторого контекста. При этом полученное множество типичных существительных может содержать много «шума».

Альтернативная гипотеза заключается в том, что выделение типичных участников может основываться на информации о синтаксических связях глагола вида 'глагол–существительное'. Эта гипотеза базируется на представлении о том, что основные участники события являются актантами, а в некоторых случаях и сирконстантами соответствующего глагола (ср., например, теорию концептуальных схем (Р. Шенк, Р. Абельсон [1]), фреймовую теорию Филлмора [2]). Здесь и далее для характеристики синтаксических связей используется синтаксическое представление в терминах грамматики зависимостей. Синтаксически-ориентированный подход, предположительно, должен давать лучшие результаты. С одной стороны, среди существительных, расположенных близко к глаголу, учитываются только синтаксически связанные с ним, с другой, в кандидаты попадают связанные с глаголом существительные, находящиеся от него на большом расстоянии.

Проведенные нами серии экспериментов показали, что предположение о том, что синтаксически-ориентированный подход должен дать безусловно лучшие результаты, не оправдалось. Ниже остановимся на описании и анализе результатов эксперимента более подробно.

МЕТОДЫ ВЫДЕЛЕНИЯ КОЛЛОКАЦИЙ

Понятие коллокации

Прежде всего, необходимо определить, что представляют собой коллокации, какие методы для их выделения используются.

Во многих теоретических работах этот термин используется по отношению к несвободным устойчивым словосочетаниям. С одной стороны, они не обладают некомпозициональностью, как, например, фразеологизмы. С другой стороны, второй компонент коллокации не может быть свободно заменен на синоним (ср., например, *strong tea* vs. **powerful tea* [3], см. также [4]). При определении коллокаций для некоторой лексемы учитывается ее типичное и постоянное окружение (см., например, работы Дж. Р. Ферс [5], Х. Джексона [6] и др., а также работы Е. Г. Борисовой [7]).

Теоретические определения коллокаций, основанные на понятии контекстной предсказуемости [5], к сожалению, дают слишком размытые критерии для выделения таких единиц (например, не вполне понятно, является ли словосочетание *произнести речь* связанным). Исследование коллокаций является также актуальным направлением корпусной лексикографии. В рамках этого направления предлагается некоторая «объективизированная» процедура оценки контекстной «предсказуемости»: коллокациями считаются два или более слова, которые встретились рядом в некотором корпусе чаще чем случайно. То есть используются некоторые статистические процедуры оценки «неслучайности» того, что две лексемы оказались рядом. Кроме того, использование статистических критериев позволяет ранжировать словосочетания по степени «коллокационной» связи. Базовые методы ранжирования пар слов по степени «связанности» описаны в [8].). Среди этих методов наиболее упоминаемыми в литературе являются взаимная информация (PMI), критерий Стьюдента (T-score), мера LogLikelihood и др. (подробнее см. [8–11] и др.). Такой подход позволяет подойти к понятию «устойчивости» более гибко (обсуждение проблем традиционного лексикографического vs. статистического методов определения коллокаций см., в частности, [9, 12]).

Поскольку большинство статистических методов ранжирования чувствительны к частотности пары в исследуемом корпусе текстов, то в верху такого ранжированного списка оказываются как фразеологизированные словосочетания типа *ломать голову*, так и глаголы с их наиболее частотными и типичными актантами, такие как *ломать руку*.

В описываемом ниже эксперименте в качестве целевых коллокаций рассматриваются оба класса случаев.

Методы отбора кандидатов в коллокации

Как было сказано, необходимым этапом в процедуре выделения коллокаций является этап отбора кандидатов. Принципиальным является противопоставление двух подходов: подхода, основанного на простой контекстной близости (далее: контекстный подход), и подхода, основанного на синтаксических связях существительных с глаголом (синтаксический подход). Первый подход предполагает нахождение кандидатов в коллокаты для некоторого глагола в числе существительных, находящихся на расстоянии не более n (n слов справа и n слов слева) от этого глагола [10, 13–15]. В данном случае говорят об окне от –n до +n словоупотреблений. Поскольку нас интересуют типичные участники события, для выбора кандидатов используется также частеречный фильтр, т.е. рассматриваются только существительные (ср. также [16, 14]).

Альтернативный метод отбора кандидатов в коллокаты для некоторого глагола – это отбор на основе синтаксического анализа: во множество кандидатов попадают существительные, связанные с глаголом синтаксической связью [17–19]. В этом отношении особый интерес представляет система Sketch Engine [18, 19]. В рамках этого проекта разработана система создания лексического портрета слова на основе его лексико-грамматической сочетаемости, т.е. на основе выделения коллокатов слова с учетом определенного типа синтаксической связи. В [18] и целом ряде других работ утверждается, что список коллокатов, полученный простым контекстным методом, содержит много «шума». Опора на синтаксические связи существительного позволяет этого избежать, а также позволяет учитывать существительные, которые линейно расположены достаточно далеко от глагола [18]. При этом учет конкретных типов синтаксических связей существенным образом повышает точность.

Предыдущие исследования по извлечению V–N коллокаций

Методы выделения ‘V–N’ словосочетаний занимают особое место в исследованиях, посвященных коллокациям. Во-первых, выделение несвободных ‘V–N’ коллокаций играет важную роль при исследовании так называемых “light verb constructions”. Именно такие словосочетания представляют интерес как для лексикографов, так и в теоретическом плане. Они имеют большое значение при изучении иностранного языка, их необходимо включать в словари. Возможность автоматического или полуавтоматического выделения таких конструкций статистическими методами активно исследовалась для английского, французского, немецкого языков (например, в работах [13–15] и др.). Авторы этих работ отмечают, что привлечение информации о синтаксических связях влияет на повышение точности (см., например, [14]). Необходимо, однако, отметить, что в задачи исследователей входило извлечение несвободных конструкций с глаголами. Соответственно, интерес представлял ограниченный набор синтаксических отношений, например, отношение ‘Глагол–Объект’, как в *to make a suggestion* [15]. Выявление некомпозициональных V–N коллокаций является актуальной задачей в системах машинного перевода [20, 21].

Другое направление исследований ‘V–N’ пар – это нахождение типичных участников ситуации, в том числе существительных, которые могут занимать позицию объекта при некотором глаголе. Выделение лексем, ассоциированных с определенными глаголами, например, с глаголами *to drink* и *to eat*, позволяет получить семантические классы существительных (‘жидкости’ vs. ‘еда’, см. “What can you drink?” [10]), а также позволяет задавать сочетаемостные ограничения для некоторой лексемы или конструкции, которые помогут разрешить многозначность (см., например, [10]). Информация о типичных актантах глагола позволяет более точно извлекать модели управления глаголов, в том числе и в задаче определения семантических ролей (semantic role labeling [22]), а также при разрешении семантической неоднозначности глаголов [23].

Исследование ‘V–N’ пар статистическими методами для русского языка представлено, например, в работе [16]. В ней исследуются возможности корпуса сверхбольшого объема (более 1 млрд словоупотреблений) для составления словаря глагольной сочетаемости. Синтаксический анализ при обработке корпуса не проводился: вместо этого было сделано предположение, что группы слов, удовлетворяющие некоторым шаблонам, с большой вероятностью представляют собой синтаксически связанные сочетания (например, «следующая за единственным глаголом группа существительного синтаксически подчиняется данному глаголу»). Действительно, гипотеза подтверждается на большом объеме текстов: авторы говорят о 99% точности результата. Однако использование корпуса исключительно большого объема, как в [16], в большинстве случаев невозможно, а для менее объемных корпусов все перечисленные ранее проблемы остаются актуальными.

В рамках задачи, поставленной в настоящей работе, нас интересуют любые существительные, семантически связанные с глаголом. В частности, при использовании синтаксического метода отбора коллокатов именные коллокаты не дифференцируются по типу синтаксической связи, учитываются все существительные, синтаксически связанные с глаголом.

ЭКСПЕРИМЕНТ

Постановка задачи

Как уже отмечалось, целью эксперимента является сравнение коллокаций 'глагол–существительное', извлеченных из большого корпуса с использованием синтаксического метода и контекстного метода.

Корпус

Для получения достаточно надежных статистических данных в исследовании использовался корпус достаточно большого объема – приблизительно 9 млн словоупотреблений[1]. Корпус состоит из случайных предложений, извлеченных из различных новостных статей, опубликованных в период с апреля 2011 по апрель 2012 года.

Поскольку корпус охватывает достаточно компактный период времени и представляет собой тексты о событиях, произошедших в этот период, то он не является сбалансированным и дает картину, несколько смещенную относительно употребления тех или иных лексем. Тем не менее, разумно предположить, что при сравнении двух методов на одном и том же материале такой смещенностью можно пренебречь.

Предварительная обработка корпуса

Для обработки корпуса использовался набор инструментов, разработанных С. Шаровым И. Нивром [24]. Токенизация и морфологический анализ производился с использованием инструмента TreeTagger [25]. Параметры для русского языка были получены Шаровым на основе подкорпуса Национального корпуса русского языка со снятой омонимией. Была также произведена лемматизация – лемматизатором на основе CSTLemma [26]. Синтаксический анализ в формализме грамматики зависимостей производился парсером для MaltParser [27], обученным на корпусе SynTagRus [28].

Обработанные тексты были помещены в реляционную базу и проиндексированы. В итоговой базе содержатся словоформы; каждой словоформе приписан набор морфологических характеристик и лемма, и для каждого предложения представлен набор отношений-зависимостей. Шаров и Нивр сообщают о 95–97% точности частеречной разметки [24]. Синтаксический анализатор также демонстрирует достаточно высокую точность. По результатам внеконкурсного участия парсера в Форуме по оценке работы систем автоматического синтаксического анализа, точность синтаксического анализа по неразмеченным связям составила 91%.

В эксперименте не учитывались типы синтаксической связи, поскольку отсутствуют данные о точности маркирования синтаксических связей. Представляется, что большая часть ошибок компенсируются большим объемом анализируемого корпуса.

Выбор экспериментальных глаголов

В эксперименте рассматривались не все глаголы, встретившиеся в корпусе. Во-первых, рассматривалась сочетаемость только финитных форм глагола. Это связано с тем, что финитные и нефинитные формы имеют разные наборы активных и пассивных синтаксических валентностей: так, например, в причастном обороте существительное может выступать вершиной, а не зависимым, как при финитной форме глагола. В финитной предикации агенс, выраженный именной группой, – зависимое от глагола, в причастном обороте агенс становится вершиной, в деепричастном обороте агенс выражен нулем, именная группа, соответствующая агенсу, оказывается в другой предикации (сравним, например: (а) *принятое <– решение*; (б) *суд <– принял решение*; (с) *приняв решение, суд* ...). К тому же нефинитные формы глагола нередко омонимичны отглагольным прилагательным или субстантивированным причастиям/прилагательным: например, *данные*, *арестованный*.

Во-вторых, рассматривались только относительно частотные глаголы, чья абсолютная частота превышает в корпусе 100 словоупотреблений. Таких глаголов на корпус объемом 10 млн оказалось около 500.

Процедура извлечения коллокаций

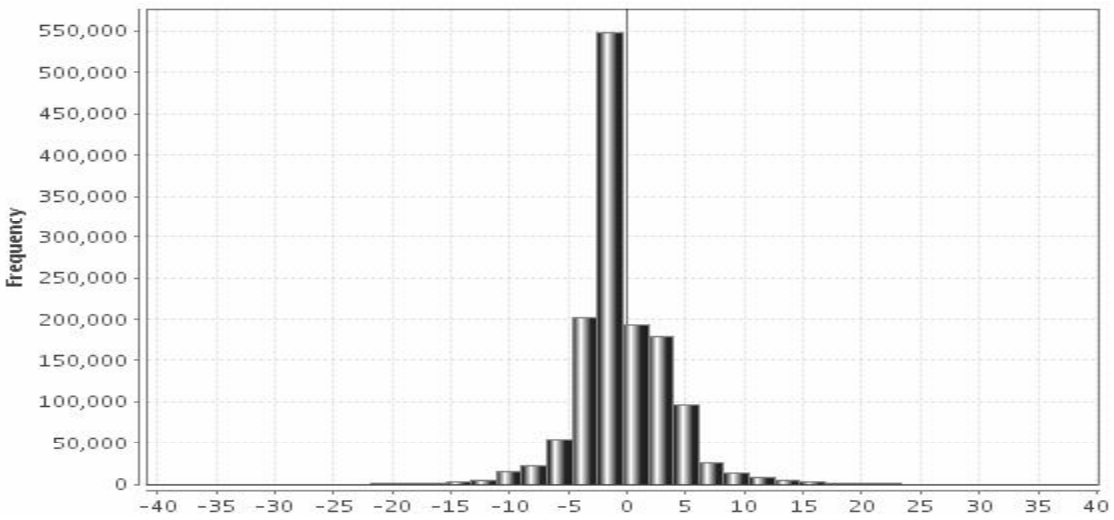
Как уже говорилось выше, коллокации извлекались из синтаксически размеченного корпуса с использованием двух разных стратегий формирования исходных списков кандидатов. Нас интересовали только существительные независимо от конкретного типа синтаксической связи и падежного оформления.

Первая стратегия состояла в том, чтобы формировать потенциальные пары коллокатов, извлекая зависимости 'глагол–существительное' безотносительно к типу синтаксической связи. Учитывались также актанты и сирконстанты глагола, оформленные предложными группами. То есть в случае управления предлогом он «пропускался» и в кандидаты попадало соответствующее существительное: так, кандидатом для глагола держать является пара 'держать–рука' в примере (3). Всего было извлечено 358 915 'V–N' пар. Далее мы будем называть коллокации, извлеченные с использованием синтаксической информации, синтаксическими коллокациями (3).

Вторая стратегия заключалась в использовании окна заданной длины. Чтобы выбрать подходящий размер окна, было рассчитано распределение расстояний между глаголами и их зависимыми существительными. Распределение расстояний представлено на графике, который наглядно показывает, что при расстоянии больше 5 в обе стороны количество существительных, связанных с глаголом, резко падает.

[1] Корпус был собран H. Christensen и доступен по адресу http://corpora.heliohost.org.

(3) В руках участники акции держали плакаты " Народу нужна справедливость , а не фронт бюрократов " .



Распределение расстояния от глагола до его аргументов

Таким образом, было решено ограничить длину окна пятью словоупотреблениями справа и пятью словоупотреблениями слева. Не-слова, например, знаки препинания и числа, игнорировались. Для всех извлеченных пар были сформированы списки коллокационных кандидатов, состоящие из леммы глагола, леммы существительного и частоты совместной встречаемости коллокации. Методом окна была извлечена 708 131 пара коллокаций. Далее мы будем называть коллокации, полученные этим методом, **контекстными коллокациями**.

Кандидаты на коллокации ранжировались с помощью метрики PMI, которая рассчитывается по формуле 1:

$$PMI(x,y) = \log \frac{P(x,y)}{P(x) * P(y)}$$

Формула 1. Pointwise mutual information

Фрагмент верхней части списка коллокаций приведен в табл.1.

Таблица 1

Значения PMI для синтаксических пар 'глагол–существительное'

verb	*noun*	*PMI*
произвести	фурор	14.6247
сойти	конвейер	13.4103
нанести	урон	13.2984
внести	лепта	13.2297
сойти	рельс	13.2050
удовлетворить	ходатайство	12.9685
пропасть	весть	12.9394
внести	корректива	12.9366
потерпеть	крушение	12.8547
выдать	ордер	12.6185
потерпеть	поражение	12.5566

PMI как статистическая мера связанности слов имеет несколько недостатков, в частности, она переоценивает значимость редких словосочетаний ([29] и др.). С этим недостатком обычно справляются, используя пороги отсечения по частоте. Однако, если порог частоты излишне завышен, то в коллокационные кандидаты попадет слишком мало пар, а если он слишком низкий, то возникнет много «шума». Поскольку заранее трудно определить, какой порог по частоте оптимален, была проведена серия экспериментов с различным порогом отсечения пар по частоте: 10, 5 и 2. Зависимость состава кандидатов от порога отсечения проиллюстрирована примером (4).

(4) сломать

c10wc10 syntax, **window**: *рука, нога*

c5wc5 syntax, window: *нога, нос, ребро, рука*

c2wc2 syntax: *нога, нос, ребро, результат, рука, челюсть*

c2wc2 window: *андрей, бедро, год, женщина, камера, лицо, мальчик, матч, нога,*

нос, падение, палец, побои, раз, ребро, результат, рука, челюсть, шея

Нотация *Cn* используется для обозначения порога совместной встречаемости величиной *n* в синтаксической модели, а *WCn* – для порога величиной *n* в контекстной модели. Например, *c10wc5* обозначает комбинацию порогов, при которой применялись пороги 10 и 5 для синтаксиса и окна соответственно.

Как видно из примера, порог отсечения 10 для синтаксических коллокаций оставляет только двух кандидатов – наиболее частотные существительные (ср. $fr_{нога}$=1187, $fr_{рука}$=3420 vs. $fr_{ребро}$=54). При этом существительные в списке для порога отсечения 10 и в списке для порога отсечения 5 не различаются семантическим классом.

Методы сравнения

Списки кандидатов для каждого глагола были ранжированы по PMI; для дальнейшего анализа были выбраны только 20 коллокаций из верхней части списка. Чтобы сравнить степень пересечения списков для каждого глагола, полученных двумя методами, были вычислены два значения взвешенной меры пересечения по формуле 2:

$$WI(A,B) = \frac{|x \in (A \cap B)|}{|x \in A|}$$

Формула 2. Взвешенная мера пересечения

Пусть Window – список коллокаций в пределах окна, syntax – список синтаксических коллокаций. Взвешенная мера пересечения WI(Window,Syntax) показывает, насколько список, полученный контекстным методом, включен в список, полученный синтаксическим методом. Мера WI(Syntax,Window), наоборот, отражает пропорцию слов из синтаксического списка, представленную в контекстном списке. Эти меры можно рассматривать как аналогию с традиционными оценками, принятыми в информационном поиске, – точностью (Presicion) и полнотой (Recall), при условии, что синтаксический список является эталоном. По аналогии с точностью и полнотой по стандартной формуле было рассчитано гармоническое среднее F между этими двумя переменными:

$$F_1 = \frac{2 * WI(window, syntax) * WI(syntax, window)}{WI(window, syntax) + WI(syntax, window)}$$

Формула 3. F-мера для WI

Результаты

В результате проведенных экспериментов были получены меры для степени совпадения контекстного списка с синтаксическим (WI(Window,Syntax) и степени пересечения синтаксического списка с контекстным (WI(Syntax,Window)). Сравнение результатов представлено в табл. 2.

Анализ результатов показывает невысокое пересечение списков. Наивысшее значение F_1 достигается при использовании порога, равного 10 для обоих алгоритмов (см. табл. 2). Однако при более подробном анализе результатов видно, что высокая F-мера при одинаковых порогах для синтаксических и контекстных кандидатов (10–10) достигается не столько за счет сближения результатов двух списков, сколько за счет того, что синтаксические списки во многих случаях полностью вложены в контекстные, но имеют существенно меньший объем. То есть усредненное значение WI(Window,Syntax) значимо выше, чем усредненное значение WI(Syntax,Window). Это отражает тот факт, что большинство слов, полученных при помощи синтаксического списка, включены в контекстные списки, в то время как обратное неверно. Так, например, для соотношения порогов синтаксических и контекстных списков 10–10 количество глаголов, для которых синтаксический список полностью вкладывается в контекстный, составляет почти 60% (для 247 глаголов из 418). Наивысшее значение WI(Syntax,Window) достигается при значении порога 5 для синтаксиса и значении порога 10 для оконного метода отбора кандидатов.

Таблица 2

Сравнение списков синтаксических и контекстных коллокаций

WI(window,syntax)				*WI(syntax,window)*			
	wc10	**wc5**	**wc2**		**wc10**	**wc5**	**wc2**
c10	0.621086	0.278444	0.117607	**c10**	0.937296	0.840384	0.604937
c5	0.79878	0.550382	0.20042	**c5**	0.605834	0.880747	0.641428
c2	0.678665	0.666004	0.496304	**c2**	0.228424	0.449963	0.696687
average=0.48974				*average=0.65396*			

F-measure			
	wc10	**wc5**	**wc2**
c10	0.718474	0.384284	0.174289
c5	0.663675	0.647521	0.269283
c2	0.30926	0.511223	0.555424

В соответствии с первоначальной гипотезой, использование синтаксиса должно было дать немного «шума», списки же коллокаций, полученные контекстным методом, – существенно больше и страдать от недостатка точности. Однако экспертный анализ показывает, что коллокации, извлеченные контекстным методом и отсутствующие в синтаксических списках, в значительном числе случаев отвечают поставленной задаче и могут считаться правильными.

ОБСУЖДЕНИЕ

Состав списков

Коллокации и типичные актанты.

Рассмотрим более подробно, какие существительные в результате попали в списки коллокатов для соответствующих глаголов по PMI. Примеры из табл. 1 показывают, что методы, использованные в работе, позволяют находить устойчивые несвободные словосочетания с глаголом. В таблице представлены пары ‘глагол–существительное’ с самым высоким PMI. Как видим, все 10 первых словосочетаний являются идиоматичными (*произвести фурор, сойти с конвейера* и пр.).

Как и ожидалось, в списках представлены как существительные, образующие с глаголом несвободные словосочетания, так и существительные, обозначающие типичных для данного глагола участников ситуации или обстоятельства места и времени. Рассмотрим пример (5):

(5) прочитать c10wc10
syntax: *интернет, книга* **window**: *интернет, книга, лекция*

В данном примере представлено как несвободное сочетание *прочитать лекцию*, так и типичный актант глагола *прочитать – книга*. В контекстном списке также представлен другой актант, отражающий реалии XXI века, и связанный с глаголом через предлог: *прочитать в интернете*.

Пример (6) иллюстрирует тот факт, что предложенный в работе метод позволяет выделять всех участников ситуации, а также обстоятельства. Так, в списках для глагола *выехать* представлены участники ситуации и обстоятельства места, типичные для новостных текстов и для события ‘действия внутренних войск’: агенсы – *группа, полиция, сотрудник (управления)*; средства – *автомобиль, машина*; цель – *полоса (встречная полоса), место (место происшествия), область.*

(6) выехать c5wc5
syntax: *автомобиль, группа, место, полоса, раз* **window**: *автомобиль, глава, год, группа, движение, дом, машина, место, область, полиция, полоса, происшествие, сотрудник, управление, человек*

Случаи «шума».

Как видно из примера (6), в списки могут попадать высокочастотные существительные, типичные для сирконстантных позиций, такие как *год*. Они могут встречаться практически с любым глаголом и, следовательно, не являются «типичными» для конкретного глагола участниками ситуации. Их можно отнести к классу «шума» с точки зрения поставленной в работе задачи. Данные существительные значительно чаще встречаются в контекстных списках.

Второй источник «шума» – это частотные одушевленные существительные, такие как *человек. …* Это связано с тем, что у достаточно большого процента глаголов в качестве подлежащего выступают одушевленные агенсы.

С другой стороны, некоторые классы одушевленных существительных, а именно имена собственные, попадают в списки из-за “смещенности” корпуса, связанной с его новостной тематикой, для которой характерна высокая упоминаемость первых лиц государств, некоторых политиков и т.д. (см. примеры (7) и (8)).

(7) уволить c5wc5
syntax: *работа* **window:** *год, медведев, работа, тренер*

(8) посоветовать c5wc5
syntax: врач **window:** *внимание, врач, знакомый, путин*

В то же время именованные сущности, хотя и в гораздо меньшем объеме, встречаются также и в синтаксических списках (см. пример (9)). Это свидетельствует о том, что попадание имен в список обусловлено не недостатками контекстного метода, а лексической смещенностью корпуса. Вопрос о том, как проявился бы этот эффект на более сбалансированном корпусе, остается открытым.

(9) заметить c10wc10
syntax: *александр, андрей, глава, депутат, заместитель, министр, нарушение, председатель* **window:** *александр, андрей, виктор, владимир, г-н, глава, губернатор, депутат, директор, дмитрий, женщина, заместитель, медведев, министр, нарушение, председатель, сергей, улица, эксперт*

Еще одним источником так называемого «шума» в контекстных списках являются части устойчивых словосочетаний. Так, например, существительное *движение* из (6) является частью устойчивого словосочетания – именной группы *полоса встречного движения* (ср. *выехать на полосу встречного движения*), аналогично существительное *происшествие – выехать на место происшествия*. Такой тип шума отсутствует в синтаксических списках. Для выделения компонентов некоторого события интерес представляют также устойчивые сочиненные глагольные группы, как в примере (10), где представлена статистически устойчивая цепочка событий: *не справился с управлением и выехал*. В такой ситуации в контекстные списки попадают актанты ситуации, заданные другим глаголом:

(10) 

Сравнение результатов контекстного и синтаксического методов

Соотношение синтаксических и контекстных списков.

Как показывают результаты сравнения (см. табл. 2), контекстные списки включают больше существительных, чем синтаксические. Более того, для порогов отсечения 10–10 наборы существительных, выделенных синтаксическим методом, полностью вкладываются в наборы, полученные контекстным методом, для 60% глаголов. При этом 104 глагола имеют только одно существительное в синтаксическом списке. В табл. 3 приведены фрагменты списков для порогов 10–10.

Предварительный анализ отдельных примеров, казалось бы, говорит в пользу гипотезы о том, что синтаксический метод должен быть более точным: в соответствующих списках должно оказаться меньше нерелевантных существительных, в списки должны попасть существительные, отделенные от своих вершин-глаголов придаточными предложениями и другими конструкциями.

В примере (11) пара *уменьшиться–доля* разделена придаточным предложением длиной 7 словоупотреблений. В пределах 5 словоупотреблений рядом с глаголом встречаются существительные *процесс, курс*, которые никак семантически с глаголом не связаны.

Однако более подробный анализ результатов показывает, что:

1) контекстные списки больше, чем синтаксические, не только и не столько из-за того, что в них много не связанных с глаголом существительных;

2) в синтаксических списках нередко отсутствуют существительные, образующие с глаголом несвободные словосочетания – лексические функции типа *принять решение*.

Сравним списки существительных, полученные двумя способами при пороге отсечения 5 (пример (12)). Пересечения списков выделены, релевантные коллокации из оконного списка подчеркнуты.

(12) снимать c5wc5

syntax: *год, квартира, комната, фильм*

window: *видео, видеокамера, время,* ***год, квартира****, кино,* ***комната****, оператор,* ***фильм***

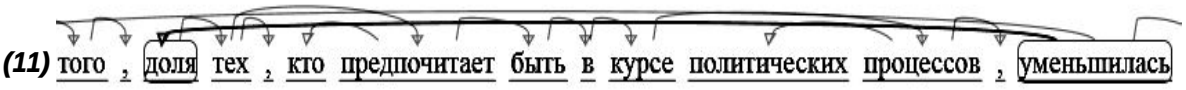


Таблица 3

Сравнение списков синтаксических и контекстных коллокаций

Глагол	*Синтаксический метод*	*Контекстный метод*
Снять	год, фильм	год фильм
Идти	Бой, борьба, война, время, год, дело, игра, место, процесс, работа, разговор, речь, человек	бой, борьба, война, время, год, город, дело, день, игра, место, процесс, путь, работа, разговор, речь, строительство, театр, улица, ход, человек
Понять	человек	время, год, деньги, жизнь, момент, человек
Продлить	арест, год, контракт, срок, суд	арест, год, контракт, срок, суд
Приобрести	год, компания, миллион, популярность, характер	акция, год, доля, компания, миллион, опыт, популярность, характер
Привести	возникновение, дефицит, изменение, итог, качество, мнение, падение, повышение, порядок, последствие, потеря, появление, пример, пример, результат, рост, рост, снижение, увеличение, удорожание	возникновение, война, дефицит, изменение, конкуренция, падение, повышение, последствие, потеря, появление, пример, рост, снижение, сокращение, тариф, топливо, увеличение, удорожание, уменьшение, цифра

Как видно из примера, контекстный список полностью включает синтаксический. В контекстном списке в число объектов, которые обычно ‘снимают’, помимо ‘фильма’, попадает ‘кино’, в списке оказываются также типичный инструмент съемки *видеокамера* и типичный агенс *оператор*. При этом существительное *год,* которое можно было бы в данном случае счесть «шумом», встречается в обоих списках. Это существительное – высокочастотное, встречается в именной группе – обстоятельстве времени, т.е. потенциально частотно при достаточно широком множестве глаголов. Это подтверждается корпусными данными: при пороге совместной частоты 10 это существительное попадает в синтаксические списки 160 из 400 глаголов и в контекстные списки 240 глаголов.

Проблемы лакун в синтаксических списках

На наш взгляд, особого внимания заслуживает вопрос, почему синтаксические списки менее полные.

Безусловно, отчасти это связано с неточностью работы синтаксического анализатора. В ряде случаев существительное «не добирает вес» для того, чтобы попасть в синтаксический список в силу того, что, например, на 10 предложений, в которых оно встречается с глаголом, в двух оно не связывается с глаголом из-за ошибок синтаксического анализа.

Однако подробный анализ примеров показывает, что достаточно большое количество расхождений в списках обусловлено с тем, что участник ситуации связан с глаголом опосредованной синтаксической связью, а не напрямую. Такое возможно в сложном предложении, когда глагол и семантически связанное с ним существительное находятся в разных предикациях или в сочинительных конструкциях, как в примере (13) (глагол в придаточном относительном, коллокат – в главном) или в примере (14) (участники ситуации входят в сочинительную группу, с глаголом связан только один). Также это возможно, если существительное является зависимым в группе, где главное слово обозначает количество (*множество **демонстрантов***) или является вершиной некоторой коллокационной конструкции ‘ход X-а’ в (15).

Еще одна ситуация, которая важна для извлечения фактов из текста, – это случаи, когда в качестве непосредственного актанта глагола выступает имя собственное или местоимение, а «типовое» существительное (играющее роль конкретного референта в ситуации) содержится в приложении, как, например, в (16).

Отметим, что в задачах извлечения фактов список таких существительных играет принципиальную роль, поскольку они могут служить маркерами ролей именованных сущностей в событии.

Еще одна причина расхождений – это то, что в контекстных списках участники ситуации, не занимающие позиции подлежащего и прямого дополнения, представлены более полно:

(17) утвердить
syntax: *депутат, правительство, программа, совет, список*
window: *бюджет, год, депутат, директор, заседание, кандидатура, компания, москва, план, правительство, президент, программа, рф, собрание, совет, список*

В примере, в контекстном списке, помимо основных участников ситуации представлено и типичное место (событие), где (в ходе которого) происходит ‘утверждение’): *собрание, заседание*.

(13) После лекции , которую он прочитал нам в школе .

(14) по их следам во Владимир выехали полицейские и сотрудники военкомата .

(15) следить c5wc5
syntax: *ход* **window:** *ход, голосование*

Как член комиссии , следил за ходом голосования на дому .

(16) ехать c5wc5
syntax: *вагон машина* **window**: *автобус, вагон, водитель, год, машина, минута, человек*

Вот он , Средний Московский Водитель , едет на своей «девятке»

Пример (17) также иллюстрирует тот факт, что контекстные списки дают более полные наборы типичных участников ситуации с некоторой ролью: например, по контекстному списку видно, что утверждают, кроме списка программы, еще и план, кандидатуру и бюджет. При этом, как уже отмечалось, «лишними» в контекстном списке оказываются лексемы, входящие в состав именных групп, которые представляют собой устойчивые словосочетания и являются актантами глаголов: *привести к сокращению доходов, бюджет на год*. Однако не всегда это действительно лишняя информация. Если исходить из задачи «собрать» типичные признаки некоторого события, то оказывается, что зависимые в именной группе, обозначающей участников ситуации, задаваемой глаголом, также играют немаловажную роль:

(18) *...приговорен к лишению свободы с отбыванием наказания в колонии строгого режима*

В (18) глагол *приговорить* имеет такую валентность, как именная группа, обозначающая участника ситуации ‘содержание приговора’, выраженного именной группой *лишение свободы с отбыванием наказания в колонии строгого режима*. С глаголом связана (через предлог) только вершина именной группы – *лишению*. Однако для полноты картины события ‘приговорить’ то, что в контекстный список попадают существительные *колония, наказание, свобода,* оказывается нелишним.

В результате, как показывает подробный анализ списков, оказывается, что контекстные списки более полные с точки зрения характеристики некоторого события. Однако они требуют дополнительного анализа для того, чтобы извлечь из такого списка характеристик именно существенные. С другой стороны, у синтаксического метода, возможно, есть перспективы, если учитывать не только ближайшие синтаксические связи, но и определенные типы опосредованных связей, не только отдельные существительные, но и типичные зависимые в именной группе, обозначающей некоторого участника ситуации.

ЗАКЛЮЧЕНИЕ

Таким образом, было проведено исследование по сравнению двух методов извлечения коллокационных пар ‘глагол–существительное’. Материалом исследования послужил корпус новостных текстов достаточно большого объема (примерно 9 млн словоупотреблений). В первом случае кандидаты на коллокации извлекались на основе контекстной близости, во втором – извлекались пары ‘глагол-существительное’, между которыми существовала синтаксическая связь. В обоих случаях использовалась мера PMI для ранжирования пар. Сравнение результатов показало, что степень совпадения списков результирующих коллокаций невысокая. Списки коллокаций, полученные синтаксическим методом, оказались неполными по сравнению с контекстными списками. Результат показывает, что для выделения типичных участников события необходимо учитывать не только непосредственные синтаксические связи существительных с глаголом, но и опосредованные синтаксические отношения. Анализ результатов, полученных контекстным методом, показал, что при достаточно большом объеме корпуса данный метод дает вполне приемлемые результаты.

СПИСОК ЛИТЕРАТУРЫ

1. Шенк Р., Абельсон Р. П. Скрипты, планы и знание // Труды IV Международной конференции по искусственному интеллекту. Т. 6. – М.: Научный совет по компл. пробл. «Кибернетика» АН СССР, 1975. – С. 208–220.
2. Филлмор Ч. Фреймы и семантика понимания // Новое в зарубежной лингвистике: Когнитивные аспекты языка. Вып. XXIII. – М.: Прогресс, 1988. – С. 52–92.
3. Halliday M. A. K. Lexis as a linguistic level // CE Bazell et al.: In memory of JR hrth. London: Longman. – 1976. – P. 150–61.
4. Виноградов В. В. Русский язык. – М.: Государственное учебно-педагогическое изд-во, 1947.
5. Firth J. R. et al. A synopsis of linguistic theory, 1930–1955 // Studies in Linguistic Analysis. Special volume of the Philological Society. – Oxford: Blackwell, 1957. –P. 1–32.
6. Jackson H. Words and their Meaning. – London and New York: Longman, 1995.
7. Борисова Е. Г. Коллокации. Что это такое и как их изучать? – М: Филология, 1995.
8. Manning M., Christopher D., Schütze H. Foundations of statistical natural language processing. – Boston: MIT Press, 1999.
9. Хохлова М. Экспериментальная проверка методов выделения коллокаций. // Slavica Helsingiensia, 34. – Helsinki: Helsinki University Press, 2008. – P. 343–357.
10. Church K. W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. – 1990. – Vol. 16(1). – P. 22–29.
11. Khokhlova M. Extracting Collocations in Russian: Statistics vs. Dictionary // JADT 2008: actes des 9es Journées Internationales d’Analyse Statistique des Données Textuelles / ed. S. Heiden , B. Pincemin. – P. 613–624.
12. Ягунова Е. В., Пивоварова Л. М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Научно-техническая информация. Сер. 2. – 2010. – № 6. – С. 30–40.
13. Breidt E. Extraction of V–N-collocations from text corpora: A feasibility study for German // Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. – Columbus, 1993. – P. 74–83.

14. Todirascu A., Tufis D., Heid U., Gledhill C., Stefanescu D., Weller M., Rousselot F. A Hybrid Approach to Extracting and Classifying Verb+Noun Constructions // Proceedings of LREC'2008, Marrakesh, Morocco. – URL: http://www.lrec-conf.org/proceedings/lrec 2008/pdf/500_paper.pdf (дата обращения: 02. 04. 2013).

15. Todirascu A., Gledhill C. Extracting Collocations in Context: The case of Verb–Noun Constructions in English and Romanian // Recherches Anglaises et Nord-Américaines (RANAM). – Strasbourg: Université Marc Bloch, – P. 107–122.

16. Клышинский Е., Кочеткова Н., Ливинов М., Максимов В. Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов // Труды международной конференции ДИАЛОГ'2010. – Бекасово, 2010. – С. 181–185.

17. Lin D. Automatic Retrieval and Clustering of Similar Words // COLING-ACL98. – Canada: Montreal, 1998.

18. Kilgarriff A., Tugwell D. Sketching words, lexicography and natural language processing: A Festschrift in Honour of B. T. S. Atkins // EURALEX / ed. Marie-Hélène Corréard. – 2002. – P. 125–137.

19. Khokhlova M. Applying Word Sketches to Russian. In: Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing. – Brno: Masaryk University, 2009. – P. 91–99.

20. Orliac B., Dillinger M. Collocation extraction for machine translation // Proceedings of Machine Translation Summit IX. – LA: New Orleans, 2003. – P. 292–298.

21. Pado S., Lapata M. Dependency-based Construction of Semantic Space Models // Computational Linguistics. – 2007. – Vol. 33(2). – P. 161–199.

22. Gildea D., Jurafsky D. Automatic Labeling of Semantic Roles // Computational Linguistics. – 2002. – Vol. 28(3).

23. Кустова Г., Толдова С. НКРЯ: семантические фильтры для разрешения многозначности глаголов // Национальный корпус русского языка 2006-2008. Новые результаты и перспективы. – Санкт-Петербург: Нестор–История, 2009.

24. Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). – М.: РГГУ, 2011. – С. 591–604.

25. Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. –Manchester. – 1994. – Vol. 12, Issue 4. – P. 44–49.

26. Jongejan B., Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – Singapore: Association for Computational Linguistics, 2009. – P. 145–153.

27. Nivre J., Hall J., Nilsson J. Maltparser: A data-driven parser-generator for dependency parsing // Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). – 2006. – P. 2216–2219.

28. Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency treebank for Russian: concept, tools, types of information // Proceedings of the 18th conference on Computational linguistics (COLING '00). – 2000. – Vol. 2. – P. 987–991.

29. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. – Toulouse, France, 2001. – P. 188–195.

Материал поступил в редакцию 19.03.13.

Сведения об авторах

АКИНИНА Юлия Сергеевна – научный сотрудник Центра Семантических Технологий НИУ ВШЭ, Москва

E-mail: jakinina@hse.ru

КУЗНЕЦОВ Илья Олегович – младший научный сотрудник Центра Семантических Технологий НИУ ВШЭ, Москва, Россия, аспирант филологического факультета ВШЭ

E-mail: iokuznetsov@hse.ru

ТОЛДОВА Светлана Юрьевна – кандидат филологических наук, доцент кафедры теоретической и прикладной лингвистики филологического факультета МГУ им. М. В. Ломоносова, старший научный сотрудник Центра Семантических Технологий НИУ ВШЭ, Москва

E-mail: toldova@yandex.ru