ОТБОР СЛОВОСОЧЕТАНИЙ ДЛЯ СЛОВАРЯ СИСТЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ

AUTOMATED ANALYSIS OF MULTIWORD EXPRESSIONS FOR COMPUTATIONAL DICTIONARIES

Лукашевич Н.В. *(louk@mail.cir.ru),* ***Добров Б.В.*** *(dobroff@mai.cir.ru),* ***Чуйко Д.С.*** *(Dasha_C@mail.ru)*
Научно-исследовательский вычислительный центр МГУ им. М.В. Ломоносова;
АНО Центр информационных исследований

В статье описываются принципы работы системы автоматизированного анализа словосочетаний, помогающей экспертам обнаруживать особенности словосочетаний на основе их компонентой структуры. Система получает на вход список словосочетаний, автоматически извлеченный по текстовой коллекции, и в качестве словарной базы анализа использует лингвистический ресурс тезаурусного типа. Эксперименты с системой мы проводим в рамках построения Онтологии по естественным наукам и технологиям - ОЕНТ.

1. Введение

Известно, что компьютерные словари, словари систем автоматической обработки текстов должны включать описания не только значений отдельных слов, но и разного рода устойчивых словосочетаний, терминов. Поэтому важным является ответ на два вопроса. Во-первых, каковы должны быть принципы отбора словосочетаний в словарь. Во-вторых, насколько процесс отбора словосочетаний может быть автоматизирован.

С одной стороны, чем больше в компьютерных словарях описано словосочетаний, тем меньше проблем с разрешением многозначности отдельных слов, тем больше будет зафиксировано специфических случаев сочетаемости. Так, в работах Большакова И.А. [2] предлагается набирать в специальную базу Кросс-лексика все встретившиеся словосочетания. С другой стороны, большие объемы неструктурированного материала трудно использовать в приложениях компьютерной обработки текстов.

Традиционным подходом является описание в компьютерных словарях семантически связанных словосочетаний (идиом, фразеологизмов), которые демонстрируют какие-либо отклонения в синтаксическом и/или семантическом поведении [1, 7]. Спектр таких устойчивых словосочетаний очень широк: от жестко фиксированных словосочетаний, которые могут рассматриваться как «слово с пробелами», до словосочетаний, которые подчиняются практически всем синтаксическим и семантическим правилам языка лишь за некоторым исключением. В последнем случае сразу обнаружить такую особенность может быть весьма сложно.

В работе Sag et.al. [12] обсуждается важный вид словосочетаний, называемых авторами институциональными выражениями. Для таких выражений характерно то, что по большей части эти выражения выглядят как свободные словосочетания, однако их компоненты не всегда могут быть заменены синонимами. Кроме того, частотность такого словосочетания очень высока по сравнению с теми словосочетаниями, которые образованы заменой слов-компонентов на синонимы. Примером таких словосочетаний является словосочетание *phone booth (телефонная будка)*. Так, и в русском, и в английском языке попытка замены слова *booth (будка)* на другие слова, например, кабина, приводит к многократному снижению частотности употребления.

Задача автоматизации отбора устойчивых словосочетаний, терминов в словарь далека от полного решения. Предложено множество методов и алгоритмов извлечения устойчивых словосочетаний, терминов из текстов (см. например, работы [3, 6, 8] и указанную в них литературу). Стандартные методы приводят, по большому счету, к одному и тому же результату. Обычно в результате программы отбора словосочетаний порождается список, упорядоченный по весу в соответствии с заложенной моделью. Верхняя часть такого списка наполнена терминами, устойчивыми словосочетаниями, которые необходимо включать в словарь прикладной системы.

Далее процент очевидно нужных для описания словосочетаний резко снижается, и наибольшую долю начинают составлять словосочетания, для которых очень трудно решить, нужно ли их описывать в словаре - для этого требуется серьезный дополнительный анализ. Поэтому методы автоматического извлечения словосочетаний терминологических словосочетаний достаточно трудно оценивать [6]. Привлеченные эксперты

часто дают очень противоречивые оценки [3]. При работе с большими предметными областями и корпусами даже лучшие методы извлечения терминологических словосочетаний показывают падение процента терминов с 90% на первой сотне до 60% на третьей тысяче списка извлеченных терминологических словосочетаний [6].

В реальной работе по формированию словарей верхняя часть (первые несколько сотен единиц) полученного списка достаточно быстро обрабатывается экспертами. Актуальной задачей является снижение трудоемкости работы экспертов при обработке сравнительно малочастотных словосочетаний.

В данной статье мы рассмотрим принципы выявления словосочетаний, необходимых для внесения в словарь компьютерной системы обработки текстов, возможные способы их формализации для автоматического выявления таких словосочетаний. Также мы опишем разрабатываемую нами систему автоматизированного отбора словосочетаний, которая должна помогать лингвистам, терминологам, экспертам выявлять особенности извлеченных словосочетаний и облегчать принятие решений по поводу их включения/невключения в словарь.

2. Критерии внесения словосочетаний в словарь

Рассмотрим, какие факторы можно учитывать, принимая решения о внесении словосочетаний в словарь.

Разработчики информационно-поисковых тезаурусов традиционно выделяют особое внимание отбору многословных терминов для включения в тезаурус.

Так, в стандартах по разработке информационно-поисковых тезаурусов [4, 9] указывает, что допускается включать словосочетания в словник, если в качестве опорного слова они содержат существительное и если выполнено одно из следующих условий:

значение словосочетания не выводится из значений его компонентов (*черный ящик*);

- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле (*торговля на вынос*);
- для данного словосочетания в словнике ИПТ существуют полные синонимы, например, *высоколиквидные акции – голубые фишки* ;
- отдельные слова словосочетания имеют слишком широкое значение;
- имеется общепринятая аббревиатура.

В работах [9, 11, 13] обсуждается совокупность принципов, которые могут служить (в сочетании) основанием для внесения словосочетания в компьютерный словарь:

- высокая частотность;
- высокая степень ассоциации, то есть более частое употребление друг с другом , чем с другими словами;
- синонимичность лексической единице (например, отдельному слову);
- значительная многозначность компонентов (*состояние дел, повестка дня*);
- словосочетание обозначает тип объекта, например, *телефонная будка, письменный стол.*

Таким образом, мы видим, что различные авторы предлагают различные критерии и соображения для включения многословных конструкций в словари компьютерных систем, что значительно затрудняет принятие решения в конкретных случаях.

3. Методы автоматического извлечения устойчивых словосочетаний

Существующие методы автоматического извлечения устойчивых словосочетаний, терминов обычно используют некоторое сочетание следующих факторов:

- частотные характеристики словосочетания (частотность по коллекции, взаимная ассоциация, вхождение в объемлющие словосочетания и т.п.);
- синтаксические ограничения: извлекаются словосочетания заданной синтаксической структуры: именные группы, глагольные группы;
- лексические фильтры, например, не извлекаются словосочетания, включающие географические названия.

Между тем есть еще ряд факторов, которые можно использовать, на следующих этапах отбора устойчивых словосочетаний.

Во-первых, часто словосочетания не извлекаются «с нуля», обычно имеется некоторый исходный ресурс, содержащий значения отдельных слов, уже включающий некоторый набор словосочетаний, и таким образом извлеченные словосочетания, можно анализировать относительно имеющегося словаря, тезауруса и т.п.

Во-вторых, извлеченные словосочетания можно сравнивать друг с другом и находить какие-либо особенности или сходство между ними.

В-третьих, можно изучать особенности словосочетания, проверяя употребление словосочетаний (исходных или специальным образом порожденных) в Интернет.

Например, в работе [12] предлагается использовать для извлечения устойчивых словосочетаний синонимы, описанные в тезаурусе WordNet. Поскольку одним из частых свойств семантически связанных словосочетаний является ограничение на замену одного из слов словосочетания синонимом, то предлагается исследовать сочетания синонимов с одними и теми же словами по корпусу, затем перепроверять в Интернет. Если разница частотностей таких словосочетаний значительна, то можно предлагать частотное словосочетание как устойчивое. Например, сравнивая употребление слов-синонимов *baggage* и *luggage* в сочетаниях с различными словами, можно обнаружить, что только *baggage* употребляется с таким прилагательным как *emotional*. Таким образом, можно предположить, что словосочетание *emotional baggage* является устойчивым. Однако в реальности ситуация осложняется тем, что у всех слов в составе исследуемых словосочетаний может быть по несколько значений, и встретившиеся словосочетания могут включать слова в разных значениях.

В работе [14] описан эксперимент по пополнению WordNet новыми словами и словосочетаниями, извлеченными из текстов путем встраивания в иерархии WordNet. Процедура реализуется за счет автоматического сопоставления сочетаемости слов и словосочетаний из WordNet с сочетаемостью неизвестных выражений. Проведенная нами проверка представленных авторами работы результатов показала, что большинство дополненных слов и словосочетаний представляют собой имена конкретных сущностей, а устойчивых словосочетаний и терминов очень мало. Отметим, что для сборки именованных объектов существуют специализированные эффективные методы.

Для исследования возможностей перечисленных выше дополнительных факторов в распознавании устойчивых словосочетаний, терминов была начата разработка автоматизированной системы анализа словосочетаний АРМ «Словосочетание».

Входом для системы анализа словосочетаний служат списки словосочетаний, автоматически извлеченных из текстовой коллекции на основе устоявшихся технологий извлечения [5, 8]. Словосочетания снабжены данными об их статистических характеристиках (частотности, взаимной ассоциации и т.п.). Состав словосочетания описывается набором слов в словарной форме, порожденных автоматически.

Словарной базой анализа словосочетаний является лингвистический ресурс тезаурусного типа, в котором синонимия слов описывается через отнесение к одной и той же единице тезауруса (дескриптору, синсету, понятию, концепту – далее концепт), разные значения слова отнесены к разным единицам тезауруса, а отношения между единицами тезауруса описаны в виде формализованных отношений. Таким образом, может использоваться тезаурус типа WordNet. Мы предполагаем использовать тезаурусы типа Тезаурус русского языка РуТез [5].

Словосочетания, по которым принято решение об их необходимости внесения в словарь, целесообразно заносить также в словарь тезаурусного вида. Таким образом, уже принятые решения смогут указывать влияние на дальнейший анализ словосочетаний.

4. Методы дополнительного анализа словосочетаний на примере пополнения Онтологии по естественным наукам и технологиям ОЕНТ

В настоящее время реализация и тестирование системы анализа словосочетаний проводится в рамках работ по развитию онтологии по естественным наукам и технологиям – ОЕНТ [5]. Онтология ОЕНТ представляет собой лингвистический ресурс для автоматической обработки текстов и содержит терминологию таких наук как математика, физика, химия, геология, биология, мы предполагаем ее бесплатное распространение для некоммерческого применения.

Начав работы над этой онтологией в 2004, мы собрали текстовые коллекции по разным естественным наукам (от 3000 до 8000 документов, от 50 до 90 Мб по каждой из наук), автоматически извлекли из них терминологические словосочетания (более 600 тысяч словосочетаний) и наиболее частотные их них (60 тысяч словосочетаний) стали одним из источников терминологии онтологии ОЕНТ.

В настоящее время величина онтологии ОЕНТ составляет 50 тысяч концептов, 135 тысяч терминов. Многие источники терминов (энциклопедии, учебники, учебные материалы) уже использованы для терминологического пополнения ОЕНТ, и существенным вопросом становится проверка полноты покрытия онтологии, пополнение относительно новыми терминами (еще не содержащимися в энциклопедических ресурсах, или не выделенными в заголовки энциклопедических статей). Поэтому мы снова обратились к исходным спискам словосочетаний как ресурсу пополнения онтологии.

Число оставшихся необработанными словосочетаний слишком велико для подробной работы. Поэтому для дальнейшего анализа извлеченных словосочетаний необходимо применение автоматизированных методов.

Рассмотрим, с чем сталкивается эксперт при анализе такого рода извлеченных списков словосочетаний, на следующем фрагменте списка словосочетаний, собранных для ОЕНТ (упорядочен по алфавиту):

вторичное зеркало
дальнейшее продвижение
изотропный источник
кристаллический агрегат
лист растения
нервное волокно
оператор проекции момента импульса
основное уравнение динамики вращательного движения
относительная внешняя система координат
параметры атомов
переходное состояние
порошкообразный алюминий
промежуточное ядро
псевдоскалярный мезон
радиоактивный ряд
степень доктора
тепловое давление
энергетическая единица
элементарная ячейка обратной решетки

Просматривая этот список, можно достаточно легко увидеть, что словосочетание *дальнейшее продвижение* – это явно не термин, поскольку не принадлежит к какой-либо конкретной области, а *нервное волокно* – это явный термин, поскольку обнаруживается в качестве заголовков статей многих словарей и энциклопедий. Для принятия решений по остальным словосочетаниям необходимо проведение дополнительных проверок.

Можно выделить следующие основные методы автоматизированного анализа новых словосочетаний.

1) Имея базовый тезаурус, можно выявлять словосочетания, которые близки по составу к словам или словосочетаниям базового тезауруса, к другим рассматриваемым словосочетаниям, то есть структурных синонимов, например, *порошкообразный алюминий – алюминиевый порошок, лист растения – листовое растение, учебный объект – учебный предмет, энергетическая единица – единица энергии*.

На первый взгляд представляется, что выявление структурных синонимов словосочетания подчеркивает тот факт, что словосочетание не является устойчивым, и не требуется описывать его в словаре. Однако приведенные выше примеры показывают, что ситуация значительно сложнее:

- для уже занесенного в словарь словосочетания может быть найден неочевидный синоним;
- может быть выявлено, что структурные синонимы, на самом деле, не являются смысловыми синонимами, так *листовое растение* – это *не растение с листьями*, а словосочетания *учебный объект* и *учебный предмет* обозначают разные сущности;
- нахождение группы структурных синонимов в извлеченных из текстов словосочетаниях обычно означает важность обозначаемой ими темы, поможет эксперту обнаружить пропущенные термины и принять решение, например, *агрегат кристаллов – кристаллический агрегат.*

При анализе структуры словосочетания могут быть также быть обнаружены словосочетания, являющиеся синонимами отдельных слов. Так, словосочетание *лист растения* является синонимом одного из значений слова *лист.* Такие, как бы тавтологичные, словосочетания могут быть обнаружены по структуре тезауруса (концепты слов словосочетания связаны между собой тезаурусными отношениями) или по толкованиям (одно слово из словосочетания встречается в толковании другого слова словосочетания**:** *лист1 – Орган воздушного питания и газообмена у растений…*).

2) На основе структуры анализа структуры словосочетаний важно выявлять словосочетания, повышающие структурированность тезауруса, представляющих собой полезные обобщения уже описанных в ресурсе сущностей. В приведенном списке таким словосочетанием является словосочетание *параметры атомов.* Введенный концепт с тем же названием будет обобщающим для таких концептов как *ЗАРЯД АТОМА, РАЗМЕР АТОМА, АТОМНЫЙ ВЕС* и др.

3) Если в ресурсе уже сформированы некоторые семантические классы слов в виде иерархической системы, то важно проверять соответствие семантических классов зависимых слов в словосочетаниях с одним и тем же главным словом. Так, если в ряду извлеченных словосочетаний *степень гидролиза, степень диссоциации, степень дифференциации, степень диффузности*, представляющих собой сочетание слова *степень* со словами, обозначающими процессы и свойства, встретится словосочетание *степень доктора*, то это словосочетание важно предъявить эксперту, поскольку значения слова *доктор* относятся к семантическому классу =человек=. Такое непохожее поведение зависимых слов в конструкциях с одними и теми же главными словами оценивается специальными весами.

4) На современном уровне развития интернет-технологий необходимым является проведение разного рода проверок употребления словосочетания в Интернет. Через глобальные поисковые сервисы можно проверить частотность употребления словосочетания (иногда относительно высокая частотность словосочетания в коллекции может оказаться случайностью [3]), а также возможно проверять, не имеет ли словосочетание каких-либо отношений, не вытекающих из его структуры.

Так, для словосочетания *вторичное зеркало* было выявлено частотное совместное употребление в текстах Интернет со словами *телескоп* и *рефлектор*, что объясняется тем, что вторичное зеркало является частью устройства телескопа-рефлектора, то есть концепт *ВТОРИЧНОЕ ЗЕРКАЛО* должен быть включен в онтологию ОЕНТ.

Анализ частот совместной встречаемости в поисковой выдаче слов со словосочетанием *радиоактивный ряд* показывает высокую частоту встречаемости слова *семейство*. Экспертная проверка показала, что словосочетание *радиоактивный ряд* является важным термином *(цепочка радиоактивных превращений),* а словосочетание *радиоактивное семейство* приводится как синоним данного термина.

Таким образом, описанные выше процедуры позволяют обратить внимание экспертов на значимые особенности словосочетаний и легче принимать решения об их включении/невключении в онтологию.

Для первого эксперимента по автоматизации дополнительных проверок извлеченных из текстов словосочетаний мы выделили следующие по частотности 20 тысяч словосочетаний (то есть с 60-й по 80-ю тысячу извлеченных словосочетаний), и применили к ним и к онтологии ОЕНТ, как словарной базе, описанные выше процедуры.

В результате было выделено около 1000 словосочетаний (5%) с особенными характеристиками. Эти словосочетания были предъявлены экспертам вместе с указанием типа найденной особенности, что позволяет относительно легко принимать решение. На текущий момент работа системы уже привела к пополнению онтологии ОЕНТ 150 терминами-синонимами, было введено 10 новых концептов.

Мы планируем расширить анализируемую базу словосочетаний до сотен тысяч, одновременно увеличить количество фильтрующих процедур, учитывающих иерархию тезаурусных отношений и статистику встречаемости словосочетаний и их структурных компонент в Интернет.

Заключение

При создании компьютерных словарей одной из трудоемких задач является обнаружение устойчивых словосочетаний и терминов для описания их в словаре. Особую сложность представляет собой анализ большого количества средне- и малочастотных словосочетаний широкой предметной области. Проблема связана с тем, что имеется множество разных видов словосочетаний, имеющих синтаксические и семантические особенности употребления, которые не следуют из состава этих словосочетаний.

В статье мы описали принципы реализации системы автоматизированного анализа словосочетаний, помогающей экспертам обнаруживать особенности словосочетаний на основе их компонентой структуры. Система получает на вход список словосочетаний, автоматически извлеченный по текстовой коллекции, и в качестве словарной базы анализа использует лингвистический ресурс тезаурусного типа. Первые эксперименты с системой мы проводим в рамках построения Онтологии по естественным наукам и технологиям - ОЕНТ.

Список литературы

1. Баранов А.Н., Добровольский Д.О. К универсальному определению идиомы // Макет словарной статьи для Автоматизированного толково-идеографического словаря фразеологизмов. М.: Ин-т русского языка, 1991. С. 7-17.

2. Большаков И.А. Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ сер. 2. 1994. - N1. - С.11 -23.

3. Браславский П.И., Соколов Е.А. Автоматическое извлечение терминологии с использованием поисковых машин Интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая - 3 июня 2007 г.). - М.: Изд-во РГГУ, 2007. 658 с.

4. ГОСТ.7.25.2001. Тезаурус информационно-поисковый одноязычный: Правила разработки: структура, состав и форма представления: Межгосударственный стандарт 7.25. - Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2001.

5. Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н., Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL'2005)– Ярославль: ЯрГУ им.П.Г.Демидова, 2005. – С.70-79.

6. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции. – 2003, с. 201-210.

7. Добровольский Д.О. Зависит ли синтаксическое поведение идиом от их семантики // Компьютерная лингвистика и интеллектуальные технологии // Труды международной конференции «Диалог 2005». - М.: Изд-во РГГУ, 2005.

8. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.

9. ANSI/NISO. Guidelines for the Construction, Format, and Management Monolingual Thesauri. - 2003.

10. Bentivogli L., Pianta E. Extending WordNet with Syntagmatic Information // Proceedings of International Wordnet Conference (GWC - 2004). - 2004. - pp. 47-53.

11. Calzolari N., Fillmore Ch., Grishman R, Ide N., Lenci A., MacLeod C., Zampolli A. Towards Best Practice for Multiword Expressions in Computational Lexicons // Proceedings of LREC - 2002, pp.1934-1940

12. Pearce D. Synonymy in collocation extraction // Proceedings of the NAACL'01 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations – 2001.

13. Sag I., Baldwin T., Bond F., Copestake A., Flickinger D. Multiword expressions: A Pain in the Neck for NLP // Proceedings of CICLING 2002, Mexico city, Mexico. – 2002.

14.Snow R., Jurafsky D., Ng A.Y. Semantic taxonomy induction from heterogenous evidence // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL. - Sydney, Australia. – 2006. pp.801-808.