Применение классического словаря к задаче автоматизированного лингвистического анализа текста

А.А. Бабанов

Санкт-Петербургский государственный университет
anton.babanov@gmail.com

Аннотация

Статья посвящена рассмотрению метода автоматизированного анализа словарных дефиниций, дающему возможность применять классический словарь для автоматического лингвистического анализа текста. Метод базируется на конвертации словаря в электронный вид с дальнейшим построением сетевой структуры на базе его статей. Для соотнесения словарных единиц со словоформами текста используется модифицированный алгоритм Леска.

1. Роль классического словаря в процессе автоматического анализа текста

Работа с огромными массивами текстов в настоящее время является повседневной необходимостью для любого специалиста. Для каждой области знания характерна своя терминология, которая систематизируется в специализированных терминологических словарях, необходимых для фиксирования устоявшейся терминологической базы области. Особенно ощутимо это в областях науки, выделившихся сравнительно недавно. Примерами таких областей могут быть кибернетика, корпусная и математическая лингвистика и многие другие дисциплины, связанные с применением вычислительной техники к решению прикладных и научных задач.

Задачи автоматической обработки текстов обычно относят к классу Information Extraction или Data Mining. Частными примерами задач этих классов могут служить: задача автоматического реферирования, извлечения фактов, установление тональности текста, и т. п.

Глубина разметки исходных даных может быть различной, в зависимости от применяемого метода извлечения информации, но в любом случае естественноязыковой текст должен пройти ряд базовых этапов – членение текста с различной степенью подробности (возможны разные варианты, с учетом абзацев или без, с установлением границ предложений и без и т.д.) и морфологический анализ (тоже может быть различным).

Важным этапом разметки является этап снятия лексической неоднозначности. На этом этапе каждой значимой единице текста должна быть придана некоторая интерпретация, идентификатор, метка, которая бы достаточно точно указывала, в каком именно значении была употреблена данная единица автором текста.

Сама по себе многозначность может носить различный характер — в одном случае она может быть вызвана совпадением грамматических форм единицы, например существительного в именительном и винительном падеже, в другом же совпадением написания двух языковых единиц имеющих разное смысловое значение (лук как оружие и лук как растение). Естественно что снятие обоих видов многозначности важно при анализе текста, т. к. оба этих вида осложняют саму процедуру автоматического анализа, в первом случае давая дополнительные возможности построения неправильной синтаксической структуры, а во втором дополняя текст значениями потенциально не заложенными в него автором, либо затрудняя понимание текста. Первый вид многозначности обычно называют морфологической неоднозначностью, а второй — лексической.

Процесс автоматической разметки текста со снятием морфологической неоднозначности может быть произведен с помощью программных модулей, имеющих один из двух основных принципов работы – на основе правил без словаря или с использованием словарной базы.

На основе опыта эксплуатации тех и других модулей был сделан вывод, что точность работы выше у модулей, обладающих словарем. Системы, основанные только на правилах, потенциально способны обработать любое слово, т.е. покрытие ими текста можно считать близким 100%, но ошибки обобщенного алгоритма достаточно часты, т.к. в любом живом естественном языке достаточно много исключений из общих правил.

Модули, обладающие словарной поддержкой, обычно имеют меньшую степень покрытия, т.к. словарь может не содержать в себе слов, присутствующих в тексте, но для слов, которые в словаре присутствуют, результат всегда точен (предполагается что словарь не содержит ошибок, на практике подобное встречается редко, но для сравнения принципиальных достоинств и недостатков двух подходов это допущение возможно).

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

Существуют также и системы, совмещающие два этих подхода. Подобные системы обладают высокой точностью результата для слов, входящих в их словарь, и кроме этого часто удачно «предсказывают» морфологические характеристики для слов, которых в словаре нет.

Очевидно, что подобные гибридные системы сочетают в себе достоинства обоих подходов – покрытие как у систем, основанных на правилах, и точность как у систем со словарной поддержкой (для слов, входящих в словарь системы).

Процесс автоматической разметки текста со снятием лексической неоднозначности может быть произведен с помощью программных модулей, опирающихся в процессе своей работы на контекст употребления или на статистику.

Статистический подход дает положительные результаты, прост в реализации, но подходит далеко не всегда, так например интерпретация слова «для» возможна как деепричастия, и как предлога, и вариант употребления как деепричастия встречается значительно реже чем предлога, поэтому система в большинстве случаев верно выберет вариант интерпритации как предлога. Но при попытке подобным образом снять многозначность слова «кольцо» становится очевидно, что статистика его употребления в текстах в значении ювелирного изделия, либо математического термина сильно зависит от тематической подборки этих самых текстов. Отсюда следует, что для снятия лексической неоднозначности в подобных случаях понадобится предварительная дополнительная обработка исходных данных, включающая в себя в том числе классификацию текстов, что является дополнительной сложностью и существенно затрудняет сам процесс автоматического анализа.

Контекст может учитываться программными модулями двумя основными способами – с помощью задания правил возможных (или не возможных) последовательностей употребления единиц текста или за счет содержательного анализа контекста употребления словоформы (например применяя предметную онтологию подходящей области). Подобный содержательный анализ потенциально дает более точные результаты, т.к. учитывает специфику конкретной ситуации, в которой словоформа была встречена, но с другой стороны подобный анализ определенно требует иточник данных словарного типа, который содержал бы в себе необходимую для анализа информацию.

Словарная поддержка позволяет модулям автоматической разметки текста показывать лучшие результаты, в связи с этим в настоящее время огромное внимание уделяется развитию сложных словарей-онтологий, которые будучи адаптированными для использования в автоматизированных программных модулях позволяют осуществлять разметку текста.

2. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний

Хорошим примером применения алгоритма, основанного на сложном словаре-тезаурусе, является метод снятия лексической неоднозначности, разработанный коллективом исследователей Научно-исследовательского вычислительного центра МГУ имени М.В. Ломоносова [2]. В своей работе авторы использовали тезаурус русского языка РуТез, в котором существуют два типа помет для представления неоднозначной лексики – М-неоднозначность и А-неоднозначность.

Помета М-неоднозначности ставится в случае существования в тезаурусе для одного текстового входа нескольких различных понятий. Например вход «пилот» может быть отнесен как к понятию «летчик», так и к понятию «водитель гоночного автомобиля».

Помета А-неоднозначности ставится в том случае, если слово потенциально неоднозначно, но в тезаурус включено только одно из возможных значений.

Алгоритм во время своего функционирования производит вычисление семантической близости между различными возможными понятиями тезауруса, отнесенными к встреченной в тексте словоформе, и окружающими эту словоформу понятиями. При этом он опирается на два типа контекстов – локальный и глобальный.

Локальный контекст исследовался как фиксированная линейная окрестность неоднозначного слова. Подобная окрестность часто называется окном, а количество элементов, включаемый в эту окрестность — шагом окна. Шаг окна измерялся в элементах, найденных в тезаурусе (т.е. словоформы, которым не соответствовал ни один текстовый вход тезауруса, не учитывались).

Кроме того, при анализе локального контекста были опробованы два режима – статический и динамический.

При статическом режиме использовалось жесткое изначально заданное количество элементов тезауруса до и после неоднозначного слова.

При динамическом режиме изначально брался фиксированный интервал, а в случае если с его использованием не удавалось снять неоднозначность целевой единицы, то размер интервала наращивался на N единиц в обе стороны.

При добавлении в алгоритм учета глобального контекста перед авторами встал вопрос о размере этого контекста, т.к. не очевидно, правильно ли учитывать тезаурусные единицы всего текста при достаточном размере последнего. Данную проблему авторы решили путем балансировки учета глобального и локального контекстов.

Сама по себе близость двух понятий в тезаурусе оценивалось по особенностям пути, связывающем эти два понятия. Т.к. тезаурус по своей структуре может быть рассмотрен как граф, обладающий

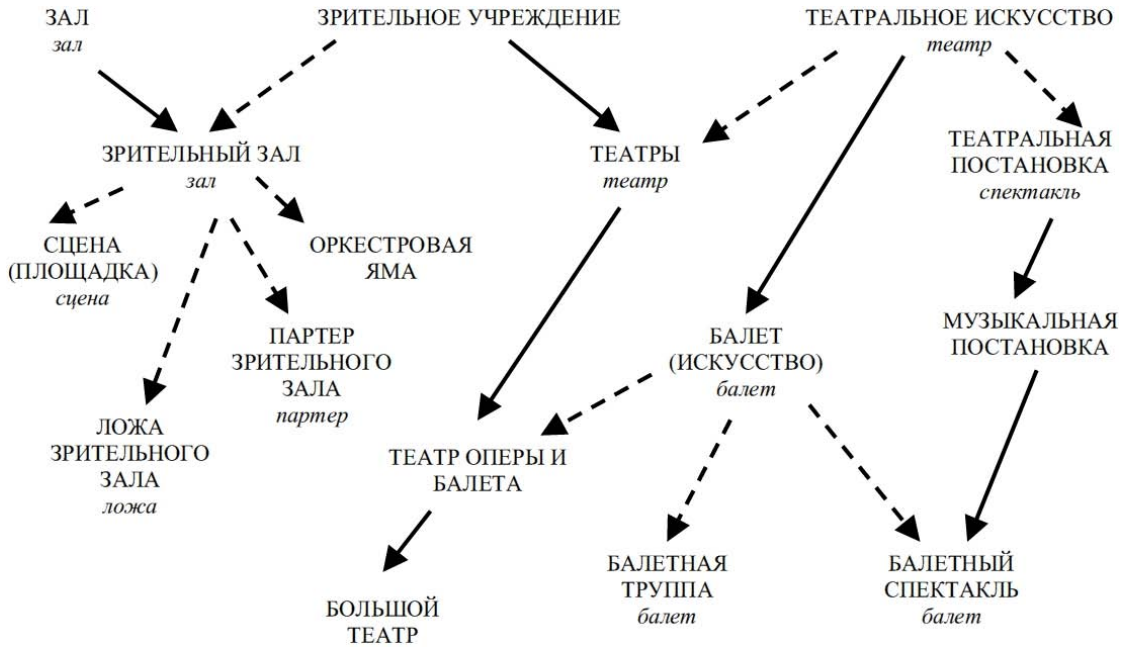

Рис.1. Фрагмент тезауруса

свойством связности, и от любой его единицы всегда есть путь до любой другой его единицы, то на путь которым могут быть связаны два понятия накладываются следующие ограничения: путь рассматривается алгоритмом только в том случае, если он состоит только из совокупности иерархических отношений, направленных в одно сторону (возможны два варианта - либо каждое следующее понятие связующего пути расположено строго вверх относительно предыдущего, либо каждое следующее понятие связующего пути расположено строго вниз относительно предыдущего), или если он содержит ровно один перегиб. Сам по себе алгоритм различает два типа перегибов – перегиб сверху и перегиб снизу. Авторы комментируют эти два вида перегибов следующим образом:

«В тезаурусе РуТез имеется три вида иерархических отношений ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация АСЦ1-АСЦ2. Таким образом три отношения (ВЫШЕ, ЦЕЛОЕ, АСЦ1) направлены по иерархии вверх, а три отношения (НИЖЕ, ЧАСТЬ, АСЦ2) – по иерархии вниз.

Для родовидового отношения ВЫШЕ-НИЖЕ определены свойства транзитивности и наследования, отношение ЧАСТЬ-ЦЕЛОЕ также рассматривается как транзитивное отношение.

На рис.1 примером иерархического пути является путь
БОЛЬШОЙ ТЕАТР:
--(ВЫШЕ) – ТЕАТР ОПЕРЫ И БАЛЕТА
--(ЦЕЛОЕ) – БАЛЕТ (ИСКУССТВО),
примером пути с перегибом сверху является путь
ОРКЕСТРОВАЯ ЯМА:
--(ЦЕЛОЕ) – ЗРИТЕЛЬНЫЙ ЗАЛ
--(ЧАСТЬ) – ПАРТЕР ЗРИТЕЛЬНОГО ЗАЛА,
примером пути с перегибом снизу является путь
ТЕАТРАЛЬНАЯ ПОСТАНОВКА:
--(НИЖЕ) – БАЛЕТНЫЙ СПЕКТАКЛЬ
--(ВЫШЕ) – МУЗЫКАЛЬНАЯ ПОСТАНОВКА»[4, с. 112]

Для проведения операции снятия лексической неоднозначности (выбора одной из возможных трактовок словоформы текста) авторы использовали числовую характеристику, приписываемую каждому значению неоднозначной словоформы. Данная характеристика являлась показателем семантической близости значения одной словоформы к значению другой. Характеристика зависела от следующих факторов:

1) чем длиннее путь по дереву тезауруса, тем слабее связь;
2) наличие перегиба делает связь слабее;
3) разные типы перегибов имеют разное влияние на характеристику;
4) наличие перегиба на высоком уровне иерархии дает больший штраф к характеристике, чем наличие перегиба на низком уровне.

Таким образом характеристика вычисляется следующим способом:
Sim(C1, C2) = максимальный_балл –длина_пути – цена_многозначности – цена_перегиба – цена_глобальности.

Максимальный балл — это наибольшее значение характеристики, применяемое в рамках системы, на момент проведения описываемого исследования он равнялся 10.

Цена глобальности — это модификатор, принимающий значения больше 0 в случае если формула применяется для работы с глобальным контекстом. При оценке локального контекста по этой формуле, данный модификатор равен 0.

Цена многозначности — это модификатор, учитывающий тот факт, что подтверждение от лексической единицы, которая в свою очередь сама является многозначной, возможно должно быть слабее. Например в фрагменте текста «светила другая, куда

более загадочная звезда» при отсутствии данного модификатора нахождение рядом слов «светила» и «звезда» приводило к трактовке обоих как небесных тел.

Данные модификаторы были подобраны эмпирически.

Основные этапы работы алгоритма следующие:

1) входной текст подвергается графематическому и морфологическому анализу;
2) на основании полученных результатов от двух последовательных анализов (цепочки лемм) происходит сопоставление каждой леммы с тезаурусной единицей и присвоение помет А-многозначности и М-многозначности;
3) проводится анализ глобального контекста, для каждой неоднозначной единицы проверяется, есть ли в тексте понятия, семантическая близость которых к ней, вычисленная по формуле указанной выше, больше 0, результаты записываются для каждой неоднозначной единицы;
4) проводится анализ локального контекста (отрезка текста заданной длины) каждой неоднозначное единицы, проверяется есть ли на указанном отрезке понятия, семантическая близость которых к выбранной неоднозначной единице больше 0, результаты данного этапа суммируются с результатами предыдущего;
5) каждому виду многозначности соответствует свое пороговое значение суммы, вычисленной на предыдущем этапе; если А-многозначная единица получила балл, превышающий пороговое значение, то считается что получено достаточное подтверждение употребления именно этого значения и многозначность считается разрешенной, если же балл ниже порогового – многозначность остается не снятой; при анализе результатов для М-многозначных единиц предпочтение отдается значению, набравшему наибольший балл.

Для оценки качества работы алгоритма была произведена разметка части корпуса вручную. Группой экспертов были создан эталонный вариант разметки, который и был после этого автоматически сопоставлен с результатами работы алгоритма. Варианты сопоставления результатов были следующие:

«

1) значение было выбрано правильно;
2) значение не было выбрано, и это было правильно;
3) значение было выбрано неправильно;
4) значение не было выбрано, и это было неправильно;
5) система выбрала один из правильных вариантов

»[3, с. 114], правильными ответами считались случаи 1,2 и 5.

Основной проблемой, проявляющейся при реализации подходов к разметке текста, основанных на применении сложных машиночитаемых словарей, является проблема высокой трудоемкости создания подобных словарей, т.к. на данном этапе построение подобных словарей подразумевает большое количество ручной работы.

В то же время подобные словари зачастую содержат взгляд на описываемую область, присущую коллективу разработчиков, а это тоже может накладывать отпечаток на результаты применения подобного словаря. Работа эксперта в данном случае может быть сложна – структуры тезаурусов подразумевают четкую концептуализацию предметной области.

С другой стороны, уже не одно столетие лексикографы составляют толковые словари, решающие задачу определения значения слова для человека. Основной проблемой их применения при автоматическом лингвистическом анализе является их человекоориентированнсть.

В связи со всем вышесказанным задачей заслуживающей внимания видится задача приведения классического человекоориентированного словаря к машиночитаемому виду, т. к. в случае создания автоматизированной системы трансформации удалось бы сэкономить большие затраты по ручному созданию автоматизированных систем словарного типа, ориентированных на поддержку автоматического лингвистического анализа.

3. Приведение классического словаря к машиночитаемому виду

Основное преимущество применения автоматизированной системы приведения классического словаря к машиночитаемому виду заключается в экономии трудозатрат, но кроме него есть и еще одно — любой словарь проходит достаточно продолжительный период опробации, т. е. признания словаря как носителя эталонных значений для набора входящей в него лексики происходит далеко не сразу после его появления. На данный момент есть классические словари, уже прошедшие подобную проверку временем, в то время как свежеразработанным автоматическим словарям-тезаурусам еще только предстоит проходить подобную проверку.

Важной проблемой приведения классического словаря к машиночитаемому виду является проблема построения системы родовидовых отношений его лексических единиц в понятных для машины терминах, которая присутствует в тезаурусах и обеспечивает необходимую поддержку процесса автоматического анализа.

Конечно, построить такую сложную систему родовидовых отношений между понятиями и создать концептуализированную базу автоматически на основе классического толкового словаря, как была показана в описанном ранее словаре-тезаурусе не представляется на данный момент возможным, но возможно создание более простой системы.

Общий алгоритм применения подобной структуры таков:

1) подбирается массив текстов обладающий жанровым и тематическим единством;

2) подбирается естественноязыковой словарь, обеспечивающий достаточное покрытие подобранных текстов;
3) словарь обрабатывается специальным программным модулем, обеспечивающим трансформацию его в структуру, удобную для использования при лингвистическом анализе текста в качестве средства снятия лексической неоднозначности;
4) проводится процедура лингвистического анализа массива текстов с применением полученной на предыдущем шаге структуры.

В классическом словаре, как и в тезаурусе, присутствует описание родовидовых отношений лексических единиц, но оно содержится в самом тексте словарной статьи каждой конкретной единицы в естественноязыковом виде. Таким образом для вычленения связей из текста статьи необходимо проводить специфический лингвистический анализ.

Установить связь между двумя взятыми значениями можно с помощью алгоритма Леска.

4. Алгоритм Леска

Алгоритм Леска это один из первых алгоритмов, разработанных для снятия лексической неоднозначности всех неоднозначных слов текста. Для его применения требуется набор словарных статей для каждого значения каждого неоднозначного слова текста и разметка отношений слов в предложениях для определения непосредственного контекста.

Основная идея алгоритма состоит в поиске наибольшего числа совпадений слов в словарных статьях значений двух неоднозначных слов.

В качестве примера применения алгоритма Леска будет рассмотрено снятие лексической неоднозначности словоформ входящих в словосочетание «pine cone» с использованием словаря The Oxford Advance Learner’s Dictionary. Для слова «pine» в словаре даются следующие значения:

1) seven kinds of evergreen tree with needle-shaped leaves;
2) pine;
3) waste away through sorrow or illness;
4) pine for something, pine to do something.

Для слова «cone» даны следующие значения:

1) solid body which narrows to a point;
2) something of this shape, whether solid or hollow;
3) fruit of certain evergreen trees(fir, pine).

Построив множество пар составленных из двух списков возможных значений и посчитав число совпадающих слов можно получить ответ, что искомыми значениями слов является первое значение для слова «pine» и третье значение для слова «cone»[1, стр.108].

5. Заключение

В заключение хотелось бы заострить внимание на следующих ключевых аспектах:

1) решение задачи автоматической лингвистической разметки текста актуально и в настоящее время;
2) конструирование для решения данной задачи сложных компьютерных словарей-тезаурусов является крайне трудоемкой задачей, с большой долей ручной работы;
3) существует метод, апробированный на материале английского языка, позволяющий применять для решения данной задачи классический толковый словарь;
4) доработка и адаптация подобного метода для применения на материале русского языка способна, если не окончательно решить, то как минимум продвинуть решение задачи автоматической разметки текста в частности в силу высокой степени автоматизации процессов, применяемых в методе.

Литература

[1] Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology), Ed. by E. Agirre, P. G. Edmonds. — 1 edition. — Springer, 2007. — November

[2] Лукашевич Н.В., Добров Б.В. Разрешение лексической многозначности на основе тезауруса предметной области. Компьютерная лингвистика и интеллектуальные технологии // Труды международной конференции "Диалог 2007" (Бекасово, 30 мая - 3 июня 2007 г.). М.: Наука, 2007. С. 400-406.

[3] Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний // Сборник "Интернет-математика — 2007", Яндекс.

[4] Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю. Многозначность как прикладная проблема: лексико-семантическая разметка в национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая — 4 июня 2006 г.) / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. М.: Изд-во РГГУ, 2006.

The use of classical dictionary to the task of automated linguistic analysis of the text

A.A. Babanov

The article is devoted to the method of computer-aided analysis of dictionary definitions, which gives the opportunity to apply the classical dictionary for automatic linguistic analysis of the text. The method is based on the conversion of the dictionary in electronic form with the further construction of network structures on the basis of its articles. To aid in matching dictionary units with the words of the text is modified Lesk algorithm