Автоматическое разрешение лексической многозначности на базе тезаурусных знаний

Н. В. Лукашевич
Научно-исследовательский
вычислительный центр МГУ
louk@mail.cir.ru

Д. С. Чуйко
Научно-исследовательский
вычислительный центр МГУ
Dasha_C@mail.ru

Аннотация

В данной работе мы описываем новый алгоритм для разрешения лексической многозначности на основе Тезауруса русского языка РуТез. Мы оценили точность работы алгоритма для задачи «все слова текста» и задачи разрешения многозначности тематической лексики.

Для задачи «все слова текста» точность применяемого метода сравнима с результатами лучших систем на специализированной конференции SENSEVAL−3, при этом мы не применяем семантически размеченные корпуса, которые используются такими системами. Однако полученная точность разрешения многозначности для всех слов текста недостаточна для применения в задачах информационного поиска.

Результаты, полученные при разрешении многозначности тематической лексики, значительно выше. Поэтому представляется перспективной разработка гибридных методов информационного поиска, сочетающих пословные методы и методы, основанные на использовании тезаурусных и онтологических ресурсов для конкретных предметных областей.

1. Введение

Одной из серьезных проблем, которые необходимо решать в рамках широкого круга систем, включающих автоматическую обработку текстов на естественном языке, является проблема автоматического разрешения лексической многозначности, то есть выбора между разными значениями слов и словосочетаний, перечисленных в лингвистическом ресурсе.

В последние годы проблема разрешения лексической многозначности стала исследоваться как отдельная задача. С 1998 г. для тестирования систем автоматического разрешения лексической многозначности проводится специальная конференция SENSEVAL (www.senseval.org).

Подходы к разрешению лексической многозначности достаточно разнообразны. Для разрешения многозначности могут использоваться некоторые внешние источники информации, например, электронные словари и тезаурусы. В качестве тезауруса обычно используется тезаурус английского языка WordNet [13]. Кроме того, для разрешения многозначности активно исследуется возможность применения методов машинного обучения, для чего обычно используются семантически размеченные корпуса. Применяются и различные комбинации отдельных методов.

Исследования методов разрешения лексической многозначности как отдельной задачи обычно делятся на два направления: разрешение лексической многозначности некоторой совокупности слов (чаще всего, несколько десятков) и разрешение лексической многозначности всех слов текста [9, 15].

В данной работе мы исследуем еще один возможный вид задачи разрешения лексической многозначности. Это задача заключается в необходимости качественного разрешения лексической многозначности для нескольких сотен − нескольких тысяч слов и связано с использованием в автоматической обработке текстов онтологий или тезаурусов в некоторой конкретной предметной области.

Мы опишем новый алгоритм по разрешению лексической многозначности слов и терминов широкой предметной области современной общественной жизни на основе Общественно-политического тезауруса. Также оценим качество работы разработанного алгоритма для задачи «все слова текста», на основе тезауруса русского языка РуТез [3], в состав которого входит Общественно-политический тезаурус.

В качестве предметной области Общественно-политического тезауруса рассматривается широкая область современных общественных отношений, проблем современного общества.

Выделение Общественно-политического тезауруса в рамках большего ресурса может быть сопоставлено с подходом, возникшим при разметке тезауруса WordNet предметными областями [11], когда дополнительно к набору тематических областей была введена специальная область Factotum. К этой области относятся синсеты WordNet, не входящие в конкретные тематические области. Именно область Factotum содержит наиболее трудно различимые значения и имеет больший процент многозначных слов [16]. Таким образом, Общественно-политический тезаурус в рамках тезауруса РуТез приблизительно соответствует объединению синсетов всех тематических областей WordNet без включения области Factotum.

Перед началом работ по разработке нового алгоритма мы оценили качество разрешения многозначности на основе разработанного в 1996 году алгоритма разрешения многозначности, использующего знания тезауруса РуТез и описанного в [2] − далее Алгоритм-96. Результаты оценки представлены в публикации [4].

2. Результаты конференции SENSEVAL

Для понимания уровня, достигнутого современными системами разрешения многозначности, важ-

но рассмотреть, каковы лучшие результаты, показанные системами на конференции SENSEVAL-3.

Для тестирования задачи «все слова текста» использовались три текста: две статьи из Wall Street Journal и фрагмент из Брауновского корпуса – общий объем 5000 слов. Всего для тестирования использовались 2081 слов. Аннотирование проводилось по набору значений тезауруса WordNet. Если в WordNet не было подходящего значения, то проставлялась пометка U.

По результатам конференции SENSEVAL-3 для английского языка в задаче разрешения многозначности для всех слов текста точность лучшей системы составляет 65,2 % [15].

Все лучшие в SENSEVAL–3 алгоритмы разрешения многозначности используют семантически размеченные корпуса по значениям WordNet. Семантическая разметка корпуса обычно используется двумя основными способами: как основа для обучения программы разрешения многозначности, и как информация о наиболее частотном значении, которое выбирается в тех случаях, когда не удалось выбрать значение с помощью основного алгоритма. По оценкам, порядка 60 % слов в тестовых текстах употреблены в наиболее частотном значении, полученному по семантически размеченному корпусу SemCor [15].

Важно отметить, что иногда в счет «благополучно» разрешенных многозначных единиц попадают также и однозначные термины. По нашей оценке, в одном из тестовых текстов около 10 % размеченных слов имеют одно значение в WordNet, например, такие слова как *congressional, constituency, salary, legislator, reelection* и др. Если рассчитать точность разрешения многозначности для лучшей системы, не считая этих однозначных слов, то величина точности разрешения многозначности лучшей системы составит 59,9 %.

Для того, чтобы изучить, насколько в приложениях информационного поиска можно использовать системы разрешения многозначности с такими показателями, в рамках конференции SemEval-2007 (http://nlp.cs.swarthmore.edu/semeval/), одним из заданий является применение алгоритмов разрешения многозначности в рамках задачи информационного поиска. Суть задания заключается в следующем: все участники должны выполнять поиск на одной и той же поисковой машине, однако перед поиском необходимо расширить запросы или тексты синонимами, соответствующими выбранным значениям.

3. Подходы к разрешению лексической многозначности на основе тезаурусных знаний

Различные алгоритмы разрешения лексической многозначности на основе тезаурусной структуры предлагались и тестировались для тезауруса английского языка WordNet.

Одним из классов предлагаемых методов является оценка семантической близости контекста вхождения того или иного многозначного термина каждому из возможных значений – синсетов.

Такая оценка близости может рассчитываться на основе сравнения путей между синсетами слов контекста и синсетами рассматриваемого многозначного слова.

В работе [10] предполагается, что два значения тем семантически ближе, чем короче связывающий их путь. Упор делается на отношения «IS-A» (является) и взвешивается длина пути относительно всей глубины таксономии (D):

$$\mathrm{Sim}_{LC}(C1, C2) = -\log (PathLen(C1,C2)/2D) \qquad (1)$$

В работе [7] предполагается что два синсета семантически близки, если соединены достаточно коротким путем, который имеет малое количество перегибов:

$$\mathrm{Sim}_{HS}(C1, C2) = c_0 - PathLen - k * d, \qquad (2)$$

где d – количество перегибов на протяжении пути; c_0 и k – константы. Если такого пути не существует, то $\mathrm{Sim}_{HS}(C1, C2) = 0$.

В экспериментах использовались значения констант $c_0 = 8$, $k = 1$, максимальная длина пути пять - шагов.

Другим направлением выбора значения многозначного слова на основе близости контекста являются подходы, основанные на оценке так называемого информационного содержания.

Ф. Резник [14] вводит характеристику «информационное содержание» (information content), которая определяется как величина вероятности встретить пример понятия С в большом корпусе $P(C)$. Эта вероятностная функция обладает следующим свойством: если $C1$ вид для С2, то $P(C1) =< P(C2)$. Значение вероятности для наиболее верхней вершины иерархии равно 1. Следуя обычной аргументации теории информации, информационное содержание понятия C может быть представлено как отрицательный логарифм этой вероятности:

$$IC(C) = -\log (P(C)). \qquad (3)$$

Чем более абстрактным является понятие, тем меньше величина его информационного содержания.

Для решения задачи разрешения лексической многозначности, вводится понятие наименьшего общего вышестоящего (LCS = Least Common Subsumer). Алгоритм базируется на идее, что нужно выбирать такое значение многозначного слова, наименьшее общее вышестоящее которого наиболее информативно.

$$\mathrm{Sim}_{Rz}(C1, C2) = IC(LCS(C1, C2)) \qquad (4)$$

Авторы работы [8] развивают формулу (4) следующим образом:

$$\mathrm{Sim}_{JC}(C1, C2) = 2 \cdot IC(LCS(C1,C2)) - \\ - (IC(C1) + IC(C2)), \qquad (5)$$

то есть учитывается не только коэффициент информационного содержания пересечения путей от синсетов, то и исходное местоположение самих исходных синсетов.

Подчеркнем, что для вычисления информационного содержания, а значит и применения описанных выше подходов необходимо иметь семантически размеченный корпус.

В работе [16] предлагается алгоритм разрешения лексической многозначности на основе разметки предметных областей Wordnet [11], при которой большинство синсетов Wordnet отнесены к той или иной предметной области, а если подходящей предметной области нет, то к специальной области Factotum.

Выбор значения многозначного слова основывается на проверке соответствия предметных областей этих значений и слов в локальном контексте (4 именные группы слева и 5 именных групп справа) и во всем тексте.

Приводятся данные, что с помощью данной системы разрешения многозначности удалось сократить количество значений на 57–65 %. При этом подчеркивается, что большинство сокращений относятся к словам из области фактотум, то есть словам, не относящимся к конкретным предметным областям таким, как *быть, начинаться, человек*.

Подход к разрешению многозначности на основе содержания целого текста тестируется в работе [6].

На первом этапе происходит сопоставление с текстом, и в специальную структуру, называемую disambiguation graph записываются все встретившиеся значения. Устанавливаются связи между узлами: гипонимы (видовые понятия), гиперонимы (родовые понятия) и понятия, имеющие с данным понятием одно и то же родовое понятие, так называемые сестры.

На втором этапе происходит разрешение многозначности в предположении «одно значение на текст».

Для каждого значения насчитывается его вес, который представляется как функция, зависящая от типа отношения и от расстояния в тексте между анализируемым вхождением и близким по смыслу значением в тексте. Так, например, синонимы, родовые и видовые значения добавляют вес к соответствующему значению, не зависимо от своего местоположения в тексте. Выбирается значение, получившее максимальный вес.

Если выбрать значение на основе полученных весов не удалось, то выбирается первое по порядку значение WordNet, которое является наиболее частотным в коллекции SemCor, семантически размеченной по значениям WordNet.

Точность разрешения многозначности на основе данного алгоритма на 35000 существительных 74 текстов Semcor оценивается как 62,09 %.

Авторы работы [12] используют алгоритм PageRank для разрешения многозначности на основе WordNet и целого текста как контекста.

Сначала для каждого значимого слова текста отмечаются все синсеты, в которые входит это слово. Такие синсеты становятся вершинами графа, ребрами графа являются отношения, полученные на основе отношений описанных в WordNet, включая:

– традиционные отношения между синсетами: гипонимия, гиперонимия, меронимия и т. п.;

– отношение номинализации, появившееся в WordNet 2.0, которое устанавливается между глаголом и существительным, являющимися дериватами;

– так называемые координатные отношения – отношения между видовыми синсетами, являющиеся подвидами одного и того же родового синсета.

Выбирается значение, получившее максимальный PageRank.

Точность разрешения многозначности данного алгоритма для задачи «все слова текста» на тестовом материале SENSEVAL–3 – 50,89 %, с учетом наиболее частотного значения – 63,27 %.

4. Тезаурус РуТез

4.1. Общая структура тезауруса РуТез

Тезаурус русского языка РуТез включает в настоящее время более 128 тысяч разных слов и выражений (с учетом разных значений более 143 тысяч), организованных в иерархическую сеть понятий.

В качестве своей подчасти тезаурус РуТез содержит Общественно-политический тезаурус (далее – Тезаурус), содержащий более 89 тысяч терминов и тематической лексики, а также набор наиболее известных географических названий (около 7 тысяч) и имен людей и организаций (около 2 тысяч).

Тезаурус РуТез обладает рядом особенностей в общей организации, в наборе отношений и др. по сравнению с WordNet [13], поэтому важно было изучить, какого качества разрешения многозначности можно достичь, используя этот ресурс.

Тестирование тезауруса РуТез как источника разрешения многозначности собственных единиц важно не только для получения результатов работы алгоритмов, но и как тестирование собственно структуры тезауруса, его состава и отношений: насколько включенные многословные конструкции помогают разрешению многозначности, насколько качественно работают цепочки иерархических отношений и др.

4.2. Представление значений терминов в Тезаурусе РуТеЗ

В Тезаурусе РуТез (далее Тезаурус) существуют два основных способа представления значений многозначных терминов.

Первым способом представления многозначности является задание одного и того же текстового входа разных понятий тезауруса (М-многозначность). Так, например, текстовый вход *пилот* сопоставлен двум разным понятиям: *ЛЕТЧИК* и *АВТОГОНЩИК*.

Второй способ представления многозначности используется в тех случаях, когда слово представлено в Тезаурусе в одном значении, но если известно, что оно может употребляться и в других значениях в целевых текстах, то ему ставится специальная пометка многозначности (А-многозначность), например.

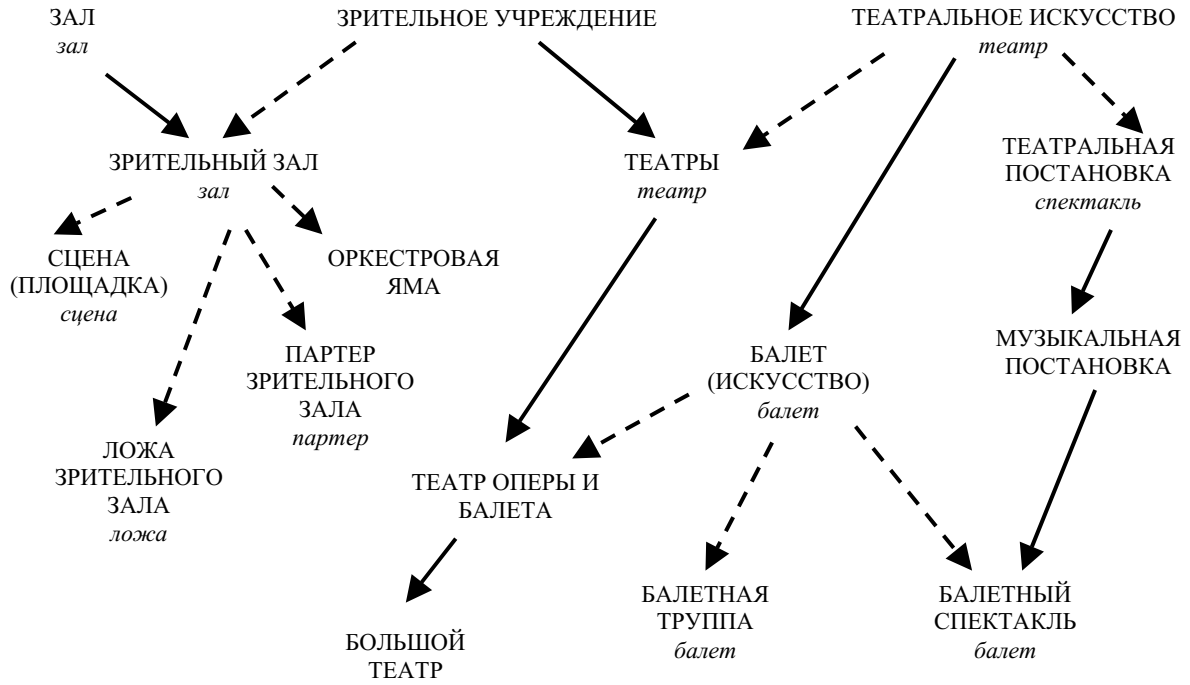

Рис. 1. Фрагмент тезаурусной сети понятий текста примера с многозначными текстовыми входами (сплошные линии – отношения ВЫШЕ–НИЖЕ, штриховые – отношения ЦЕЛОЕ–ЧАСТЬ)

Пометка многозначности часто используется для отметки географических названий, которые могут совпадать с фамилиями и именами людей, сокращениями и др., например, *Львов* (город), *Владимир* (город), *Павлово* (город в Нижегородской области).

В настоящее время тезаурус РуТез содержит более 15 тысяч многозначных единиц, из них для более 11 тысяч слов представлено несколько значений (М-многозначность), многозначность остальных отмечена пометкой.

В составе Общественно-политического тезауруса насчитывается около 6,5 тысяч многозначных терминов. Для 2204 терминов представлено два и более значений.

5. Пример

В качестве примера рассмотрим фрагмент статьи из «Независимой газеты» от 23 ноября 2003 г. под названием «Первый бриллиант Александра Волошина»:

В ***понедельник*** на ***сцене Большого театра*** ***сверкали*** «***Бриллианты американского балета***». ***Концерт*** был ***посвящен*** 70-***летию установления дипломатических отношений*** между ***Россией*** и ***США***. В ***зале сидели*** все ***мыслимые*** и ***немыслимые дипломаты*** с обеих ***сторон***. В этот ***вечер спектакль разыгрывался*** по обе ***стороны рампы***, ***точнее***, ***оркестровой ямы***. И второй, надо ***сказать***, был ничуть не менее ***захватывающим***. Пока на ***сцене звезды американского балета показывали чудеса хореографической техники***, в ***противоположной стороне партера***, в ***царской ложе***, ***светила*** другая, куда более ***загадочная звезда***.

Полужирным шрифтом выделены слова, которые включены в качестве единиц в тезаурус РуТез. Видно, что практически вся содержательная лексика включена в анализ.

Подчеркнутые слова входят в тематический подтезаурус – Общественно-политический тезаурус. Фрагмент содержит группы единиц тезауруса, относящихся к зрительному залу (рис. 1): *сцена, зал, рампа, оркестровая яма, ложа, партер*, а также к искусству*: концерт, балет, Большой театр, хореографический*, что дает возможность использования этой информации для разрешения многозначности.

Относительно Общественно-политического тезауруса фрагмент содержит 25 тезаурусных единиц, из них 15 многозначных. Такие слова, как *звезда (небесное тело), техника (техническое устройство), зал (общественное помещение), партер (зрительного зала)* представляют пример А-многозначности, то есть их другие значения не входят в состав Общественно-политического тезауруса, а многозначность отмечена только специальной пометкой.

Относительно Тезауруса РуТез все многозначные слова имеют М-многозначность, за исключением слова *партер,* другие значения которого на момент обработки еще не были описаны.

Слова, находящиеся вне Общественно-политического тезауруса (*посвящен, установления, сидели, сторон, вечер, разыгрывался, показывали* и т. д.), отчетливые смысловые группы не образуют, что может затруднить выбор правильного значения.

6. Описание алгоритма разрешения многозначности на основе тезауруса

Основой для разработанного алгоритма разрешения многозначности является оценка семантической близости между возможными значениями

и окружающим текстовым контекстом. При этом рассматривается как локальный контекст, который задается в виде некоторого окна – линейной окрестности многозначного вхождения слова, так и глобальный контекст, в который входят все слова текста.

Рассмотрение глобального контекста учитывает такое свойство связного текста как лексическую связность текста, то есть повторяемость одних и тех же лексических единиц и совокупностей семантически близких лексических единиц в связном тексте.

6.1. Учет локального и глобального контекста

В качестве локального контекста рассматривается фиксированная линейная окрестность многозначного вхождения слова, измеряемая в количестве найденных элементов тезауруса, – исследовался размер окна окрестности от 1 до 5 элементов в обе стороны.

Также мы исследовали задание локального контекста как «динамического» окна N + N, то есть сначала происходит попытка выбора значения слова в окрестности длиной N, если это удается, то обработка данного вхождения заканчивается. Если не удается, то происходит расширение окрестности еще на N элементов и процедура выбора значения продолжается. Тестировались такие динамические окна как 1 + 1, 2 + 2, 3 + 3.

При использовании глобального контекста возникает вопрос о том, насколько в достаточно длинном тексте правомерно использование полного текста как базы для выбора значения, не нужно ли вводить некоторые ограничения, например, на расстояние (в абзацах, предложениях) между данным многозначным вхождением и упоминанием семантически близкого понятия в тексте. Так, в работе [6] разные типы связи имеют разную сферу действия и разный вес в зависимости от такого рода расстояния, измеряемого в абзацах и предложениях.

В процессе экспериментов нами была выбрана следующая специфика учета глобального контекста.

В качестве элементов глобального контекста учитываются только однозначные вхождения тезаурусных единиц.

Мы не накладываем никаких ограничений на расстояние между вхождением многозначного слова и семантически близкими словами. Предполагается, что возможное неправильное подтверждение от далекой части текста должно преодолеваться правильным подтверждением от локального контекста и более близкой части текста.

Поскольку локальный контекст достаточно ограничен, а глобальный контекст может достигать весьма большой величины, то были сделаны попытки сбалансировать свидетельства в пользу того или иного значения, получаемые от локального и глобального контекстов. Прежде всего, вес подтверждения значения, получаемый от некоторой лексической единицы в локальном контексте всегда выше, чем от той же единицы, расположенной вне локального контекста. Кроме того, мы тестировали возможность применения коэффициента, уменьшающего вес подтверждения от глобального контекста при увеличении длины текста (точнее при увеличении максимальной частотности лексической единицы в тексте).

6.2. Семантическая близость понятий как функция от особенностей пути отношений между ними

Семантическая близость между двумя понятиями *C1* и *C2* оценивается на основе рассмотрения пути отношений, который существует между этими единицами тезауруса.

Между понятиями в тезаурусе могут существовать пути разной конфигурации, тезаурус связен и всегда существует путь отношений от одного произвольного понятия тезауруса до другого понятия тезауруса. Однако подобно подходу [7] мы ограничиваем конфигурации путей между понятиями *C1* и *C2*, которые рассматриваются при оценке семантической близости понятий, а именно, либо путь должен состоять из совокупности иерархических отношений, направленных в одну сторону, например, последовательность отношений от вида к роду, либо такой путь должен включать ровно один перегиб, то есть изменение направления движения. При этом рассматриваются перегибы двух видов: перегиб-сверху, например, сначала несколько отношений от видовых понятий к родовым, затем несколько отношений от родовых понятий к видовым, так и перегиб-снизу.

В тезаурусе РуТез имеется три вида иерархических отношений ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация АСЦ1-АСЦ2 [3, 4]. Таким образом, три отношения (ВЫШЕ, ЦЕЛОЕ, АСЦ1) направлены по иерархии вверх, а три отношения (НИЖЕ, ЧАСТЬ и 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*.

Построение разрешенных путей осуществляется следующим образом.

Для каждого понятия тезауруса можно определить совокупность иерархически вышестоящих понятий – так называемое «дерево-вверх». «Дерево-вверх» понятия C0 включает те понятия тезау-

руса, к которым от С0 может быть проведен путь, состоящий из отношений одной направленности, и который с помощью правил наследования и транзитивности может быть сведен к одному отношению [4]. Схожим образом, на основе иерархических отношений, направленных вниз, определяется совокупность иерархически нижестоящих понятий – «дерево-вниз».

Так, например, на рис. 1 для понятия *БОЛЬШОЙ ТЕАТР* можно видеть следующие вышестоящие по иерархии понятия (понятия из «дерева–вверх»): *ТЕАТР ОПЕРЫ И БАЛЕТА*, *ТЕАТР*, *ТЕАТРАЛЬНОЕ ИСКУССТВО*, *ЗРИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*.

Таким образом, между двумя понятиями существует путь разрешенной структуры, если либо одно из понятий входит в дерево–вниз или в дерево–вверх другого понятия, либо если между их деревьями имеется непустое пересечение.

6.3. Числовая оценка семантической близости

Семантическая близость понятий, связанных путем заданной конфигурации, зависит от особенностей пути между понятием-значением и подтверждающим понятием:

– чем длиннее путь между понятиями, тем слабее семантическая близость;

– наличие перегиба на пути ослабляет семантическую близость;

– разные типы перегибов на пути могут по-разному влиять на семантическую близость;

– перегиб пути на высоком уровне иерархии хуже, чем на более низком уровне.

Кроме того учитывался тот факт, что подтверждение от лексической единицы, которая в свою очередь многозначна, возможно должно быть слабее. Например, в тексте примера во фрагменте «*светила другая, куда более загадочная звезда*» нахождение рядом слов *светила* и *звезда,* приводит к трактовке обоих слов как небесных тел.

Для учета такого рода рассуждений была применена следующая формула:

$$\begin{aligned} \mathrm{Sim}_{new}(C1, C2) = {} & \textit{максимальный\_балл} - \\ & - \textit{длина\_пути} - \\ & - \textit{цена\_многозначности} - \\ & - \textit{цена\_перегиба} - \\ & - \textit{цена глобальности}. \end{aligned} \qquad (6)$$

Максимальный балл представляет собой максимально возможную оценку подтверждения, связанную с тем, что встретился однозначный синоним рассматриваемого многозначного термина. В настоящее время, величина максимального балла равняется 10.

Параметр *цена_глобальности* составляет величину, больше нуля, в случае оценки глобального контекста и величину, равную нулю, при анализе локального контекста.

6.4. Этапы алгоритма

Поступающий текст проходит через процедуру графематического и морфологического анализов. Далее на основе цепочек лемм, полученных в результате морфологического анализа, происходит сопоставление с тезаурусом. Для каждой сопоставившейся тезаурусной единицы отмечается ее статус: однозначное сопоставление, сопоставление с пометкой многозначности (А-многозначность), сопоставилось несколько единиц тезауруса (М-многозначность). Отметим, что если одна из сопоставленных тезаурусных единиц, полностью включается в другую тезаурусную единицу, то эта ситуация многозначной не считается, сопоставленной считается более длинная тезаурусная единица.

Процедура разрешения многозначности начинается с анализа глобального контекста. Для каждого значения неоднозначных единиц текста анализируется, упоминались ли в тексте понятия, семантическая близость которых к текущему понятию, составляет число баллов, большее 0, по формуле (6). Все набранные баллы понятий-значений многозначных единиц суммируются и запоминаются.

Далее происходит анализ локального контекста. Для каждого вхождения многозначной тезаурусной единицы просматривается заданная текстовая окрестность, выбираются упоминаемые понятия, связанные с понятиями данной многозначной единицы тезаурусными путями разрешенной конфигурации, и подсчитываются баллы по формуле (6). Баллы, полученные при глобальном и локальном анализах суммируются.

Для каждого вида многозначности задается свой порог. Если понятия-значения, получили баллы, меньшие, чем заданный порог, то считается, что ни одно значение не подтвердилось, возможно, в тексте использовано какое-то другое значение.

Если понятие единицы с А-многозначностью получает количество баллов, большее чем установленная пороговая величина, то это значение подтверждается и, соответственно, выбирается.

Среди понятий для текстовой единицы с М-многозначностью выбирается значение, получившее максимальное количество баллов.

Если понятия единицы с М-многозначностью получили одинаковое количество баллов, превышающее пороговое, то выбирается вышестоящее по иерархии понятие, так, например, для значений слова *балет* таким понятием является понятие *БАЛЕТНОЕ ИСКУССТВО* (см. рис. 1). В случае если такой иерархической связи не имеется, то в настоящее время не выбирается ни одно из понятий – многозначность остается неразрешенной. Если на основе разметки корпуса было бы известно наиболее частотное значение, то можно было в таких случаях выбирать именно это частотное значение.

6.5. Сравнение с Алгоритмом-96 обработки многозначности по Общественно-политическому тезаурусу

Прежний алгоритм разрешения многозначности, работавший на том же ресурсе, отличается по следующим параметрам.

Во-первых, в Алгоритме-96 для учета концептуальной близости использовались только пути, со-

стоящие из иерархических отношений одной направленности, то есть без перегибов, таким образом семантически близкими считались только понятия, находящиеся в иерархических отношениях между собой. Это приводило к явным проблемам на относительно коротких текстах, таких как новостные сообщения, когда необходимые для подтверждения иерархически расположенные понятия не входили в состав анализируемого текста.

Во-вторых, не было ограничений на длину пути между понятиями, что приводило, например, к тому, что многозначность очень конкретного понятия могла быть разрешена на основе нахождения в тексте очень абстрактного понятия.

В-третьих, не было весовой оценки семантической близости между понятиями на основе путей между ними или каких-либо других: подтверждение производилось на основе принципа «данет».

В-четвертых, приоритет отдавался глобальному контексту, то есть сначала проверялось, если ли подтверждение для того или иного значения по всему тексту. Если несколько значений имели подтверждение в глобальном контексте, то проверялся локальный контекст: выбиралось то значение, подтверждение для которого находилось ближе всего к исследуемому многозначному вхождению.

Таким образом, начиная работы над новым алгоритмом разрешения многозначности, мы рассчитывали на повышение качества разрешения многозначности за счет более аккуратного учета специфики путей между понятиями тезауруса.

7. Настройка и тестирование алгоритма разрешения многозначности

Для определения качества разрешения лексической многозначности необходимо было выполнить эталонную разметку найденных терминов по значениям. Для каждого документа экспертами были созданы эталонные файлы, с правильной разметкой значений.

После получения эталонных файлов они были автоматически сопоставлены с результатами работы программы разрешения многозначности. Были выделены следующие случаи соответствия (несоответствия) эталонной разметки и результирующего файла работы программы:

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.

В качестве правильных решений системы рассматривались виды соответствия 1), 2) и 5). В качестве основной характеристики работы алгоритма оценивалась точность разрешения многозначности, которая рассчитывается как отношение между числом правильных решений и числом всех решений.

Число всех решений – это количество обнаруженных в тексте единиц тезауруса, отмеченных как многозначные. Таким образом, при сопоставлении одного и того же текста с Общественно-политическим тезаурусом количество решений, которое необходимо принять, меньше, чем при сопоставлении с объемлющим тезаурусом РуТез.

Тестировались следующие параметры алгоритма:

– максимальная длина дерева, то есть насколько далеко в одном и то же направлении иерархических отношений от исходного понятия можно искать подтверждающее значение понятия – длина дерева может быть различной для локального и глобального контекстов;

– строение (статическое или динамическое см. п. 6.1) и размер окна локального контекста;

– в локальном контексте: учитывать ли в полном объеме подтверждение от многозначного термина. Если снижать вес подтверждения в таких случаях, то каким образом: вычитать баллы, делить на коэффициент и т. п.;

– цена глобальности – насколько баллы, полученные от одного и того же подтверждения, меньше в глобальном контексте, чем в локальном;

– веса различных перегибов путей для локального и глобального контекстов;

– пороги для видов многозначности: А-многозначности и М-многозначности.

7.1. Тестирование алгоритма разрешения многозначности на основе Общественно-политического тезауруса

Тестирование алгоритма разрешения многозначности для терминов Общественно-политического тезауруса проводилось на материалах газет и наборе новостных сообщений. Предварительно, случайным образом было выбрано несколько дат. Из коллекции Университетской информационной системы РОССИЯ (www.cir.ru) были выгружены газетные публикации, относящиеся к выбранным датам. Набор газетных публикаций включает полные номера газет «Известия», «Ведомости», «Независимая газета», «Комсомольская правда». Каждый номер содержит несколько десятков статей. Средний размер статьи около 5 Кб. За те же даты были взяты новостные сообщения из коллекции новостей Яндекса.

В процессе эксперимента вручную было размечено 197 документов, что соответствует полным номерам газет «Известия», «Независимая газета», «Ведомости», «Комсомольская правда» от 19 ноября 2003 года, а также было размечено 30 новостных сообщений за ту же дату. Взятие полных номеров обеспечивает достаточно большое разнообразие тематики документов.

Результаты работы алгоритма разрешения многозначности по каждому из источников показаны в табл. 1, где N_{doc} – число документов, N_{amb} – число вхождений неоднозначных терминов, P_{new} – точность по новому алгоритму, P_{96} – точность по старому алгоритму.

Совокупная точность работы системы по разра-

ботанному алгоритму (процент правильно принятых решений) в процессе тестирования составила 73,37 % и выросла на 6,7 % относительно точности разрешения многозначности, полученной по старому алгоритму.

Таблица 1

Точность разрешения лексической многозначности по источникам публикаций

Источник	N_{doc}	N_{amb}	P_{new}, %	P_{96}, %
Известия	44	2525	**75,23**	72,00
Ведомости	62	2697	**77,89**	73,41
Независимая газета	42	2776	**68,14**	66,50
Комсомольская правда	49	2240	**66,74**	63,04
Яндекс-Новости	30	450	**75,05**	68,00
Всего	227	10688	**73,37**	68,77

Как и предполагалось, наибольший рост точности удалось получить на относительно коротких текстах новостных сообщений, который составил более 10 %.

Для получения лучших результатов тестировались разные наборы параметров.

К особенностям наилучшего набора параметров можно отнести следующие закономерности.

Были выбраны разные пороги для разных видов многозначности: 4 балла для А-многозначности, и 2 балла для М-многозначности. Такой результат является предсказуемым, поскольку при М-многозначности между собой «соревнуются» несколько значений, а при А-многозначности значение-контрагент находится вне зоны тезауруса.

Выяснилось, что подтверждение от многозначного термина в локальном контексте значимо так же, как и от однозначного термина. Эта закономерность не была очевидна, при ручном анализе было видно, что между парами многозначных терминов иногда возникают ложные корреляции, приводящие к выбору неправильных значений для обоих терминов.

Наилучшей оказалась динамическая окрестность локального контекста 3 + 3.

Лучший результат был получен для высоты деревьев 2 как для локального, так и для глобального уровня, то есть при поиске семантически близких терминов в среднем лучше использовать как подтверждение понятия, отстоящие от понятий, соответствующих многозначному выражению, общая длина пути не более 4 отношений.

Из всех типов перегибов «наихудшими», получившими максимальные баллы штрафа, оказались перегибы типа: *видовое_понятие1 – родовое понятие – видовое понятие_2*, что ожидалось, а также перегиб-внизу типа: *родовое понятие_1 – видовое понятие – родовое понятие_2*.

При анализе результатов работы алгоритмов, изложенных в табл. 1, нужно подчеркнуть важное обстоятельство. Тезаурус содержит много однозначных словосочетаний, в состав которых входят многозначные слова, например, *министр обороны, уголовное дело, дополнительный отпуск*. При анализе текста эти многозначные слова попадают внутрь многословных терминов и задача разрешения их многозначности не возникает.

Если бы словосочетаний не было, то пришлось бы разрешать многозначность этих слов алгоритмически. Было подсчитано, что если учесть те многозначные слова, многозначность которых снимается за счет объемлющих словосочетаний, то точность разрешения многозначности на основе комплекса «многословные термины тезауруса + алгоритм разрешения» возросла бы в среднем на 5 %.

Также мы исследовали вопрос, насколько точность разрешения многозначности зависит от частотности многозначной единицы в тексте. Была выявлена интересная корреляция, что разрешение многозначных слов, встретившихся в тексте один раз, во всех подколлекциях на несколько процентов ниже, чем в целом по коллекции. Это означает, что точность разрешения для слов с большей частотностью выше, чем приведенная в таблице.

7.2. Тестирование алгоритма разрешения многозначности на запросах из правовой области

Исследуя эффект нового алгоритма по разрешению лексической многозначности для коротких текстов, мы сделали небольшую коллекцию 40 длинных запросов в области права из коллекции РОМИП-legal [5], например, таких как *компенсация подоходного налога при приобретении недвижимости*. Для этой коллекции разрешение многозначности терминов Общественно-политического тезауруса достигло величины 82,02 %, в то время как точность прежнего алгоритма составляла величину 48,31 %.

Для такой коллекции параметры алгоритма настраивались отдельно. Параметры, на которых были получены лучшие результаты для коллекции запросов, оказались совершенно иными, чем для коллекции статей: это максимальные величины деревьев – 7 шагов, минимальные пороги для обоих видов многозначности, минимальные цены перегибов.

Такие результаты привели к мысли, что можно сделать систему автоматической настройки параметров алгоритма в зависимости от длины обрабатываемого текста.

Был проведен следующий эксперимент: та же тестовая коллекция статей (п. 7.1) была разделена на пять подколлекций по величине текстов. Мы пытались подобрать лучшие параметры для каждой группы текстов и выявить функцию изменения основных параметров. Однако в этом эксперименте четкой корреляции, позволяющей реализовать самонастройку параметров, не было выявлено. Группа самых коротких текстов статей давала неожиданно низкий результат разрешения многозначности, причем лучший результат – 71,02 % был получен на параметрах более близких к параметрам всей коллекции, чем к лучшим параметрам, полученных для запросов.

7.3. Тестирование алгоритма разрешения многозначности для задачи «все слова текста»

Для тестирования алгоритма разрешения многозначности для задачи «все слова текста», было взято по

две статьи из газет «Известия», «Комсомольская правда», «Независимая газета», «Ведомости». Количество многозначных единиц – 1120. Меньший объем коллекции объясняется значительно большими трудозатратами по подготовке эталонной разметки. Полученная точность нового алгоритма 57,14 %, с учетом разрешения за счет попадания в словосочетания, описанные в тезаурусе – 63,4 %.

Для лучшего набора параметров этой коллекции характерна большая величина окна – используется динамическое окно 4 + 4.

Таким образом, точность разрешения многозначности, показанная реализованным алгоритмом для задачи «все слова текста», не использующая размеченного корпуса, приблизительно соответствует результатам работы лучших систем на конференции SENSEVAL.

Мы получили этот результат без использования дополнительной информации о наиболее частотных значениях, без использования размеченного корпуса и т.п. Наилучший известный авторам алгоритм, использующий только WordNet, имеет точность – 50,89 % на данных SENSEVAL-3 (вспомним еще про 10 % однозначных слов – см. п. 2).

Однако представляется, что полученные результаты точности разрешения многозначности для задачи «все слова текста» даже лучших методов недостаточны для того, чтобы использоваться в реальных приложениях информационного поиска.

8. Заключение

Разработан новый алгоритм разрешения лексической многозначности на основе тезаурусных знаний, не использующий информацию размеченных текстовых корпусов.

Для задачи «все слова текста» результаты алгоритма сопоставимы с результатами лучших систем, достигаемых комбинированными методами с использованием семантически размеченных корпусов и информации о наиболее частотном значении.

Однако достигнутой точности разрешения многозначности для всех слов текста не достаточно для того, чтобы использовать результаты разрешения многозначности в приложениях информационного поиска.

С разрешением многозначности тематической лексики ситуация принципиально другая. Достигнуты значительно более высокие результаты разрешения многозначности. Эти результаты потенциально могут быть увеличены за счет использования дополнительной информации (например, о самом частотном значении, которое можно выбирать при величинах оценки значений ниже пороговых или близких к пороговым).

Поэтому, на наш взгляд, может оказаться перспективным развитие комбинированных методов, сочетающих пословные методы обработки текстов и обработку по тематическим понятийным ресурсам таким, как тезаурусы и онтологии.

9. Благодарности

Авторы благодарят Б. В. Доброва, Т. М. Селиванову, А. В. Сидорова, М. Г. Шаталову, С. В. Штернова за вклад в проведение эксперимента.

10. Литература

[1] *Гальперин И. О.* Текст как объект лингвистического исследования / И. О. Гальперин. – М.: Наука, 1981.

[2] *Лукашевич Н. В.* Разрешение многозначности терминов в процессе автоматического индексирования /Н. В. Лукашевич // Тр. междунар. семинара Диалог'96. – М., 1996. С. 142–146.

[3] *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций / Н. В. Лукашевич, Б. В. Добров // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. семинара Диалог'2002. Под ред. А. С. Нариньяни. – М.: Наука, 2002. Т. 2. – С. 338–346.

[4] *Лукашевич Н. В., Добров Б. В.* Разрешение лексической многозначности на основе тезауруса предметной области. Компьютерная лингвистика и интеллектуальные технологии / Н. В. Лукашевич, Б. В. Добров. // Тр. междунар. конф. «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.). – М.: Наука, 2007. – С. 400–406.

[5] РОМИП. Труды третьего российского семинара РОМИП-2005. – Санкт-Петербург: НИИ Химии СпбГУ, 2005. – 226 с.

[6] *Galley M., McKeown K.* Improving word sense disambiguation in lexical chaining / M. Galley, K. McKeown // IJCAI 2003.

[7] *Hirst G., St-Onge D.* Lexical Chains as representation of context for the detection and correction malapropisms. / G. Hirst, D. St-Onge // C. Fellbaum, editor, WordNet: An electronic lexical database and some of its applications. Cambrige, MA: The MIT Press, 1997.

[8] *Jiang J., Conrath D.* Semantic similarity based on corpus statistics and lexical taxonomy / J. Jiang, D. Conrath // COLING 1997.

[9] *Kilgarriff A., Rosenzweig J.* Framework and Results for English SENSEVAL / A. Kilgarriff, J. Rosenzweig. // Computers and the Humanities, 2000. V. 34. P. 15–48.

[10] *Leacock C., Chodorow M.* Combining local context and WordNet similarity for word sense identification / C. Leacock, M. Chodorow // WordNet: An electronic lexical database. The MIT Press, 1998.

[11] *Magnini B., Cavaglia G.* Integrating Subject Field Codes into WordNet / B. Magnini, G. Cavaglia // Proceedings of the Second International Conference on Language Resources and Evaluation LREC 2000, Athens, Greece, 2002.

[12] *Mihalcea R., Tarau P., Figa E.* PageRank on Semantic Networks, with application to Word Sense Disambiguation / R. Mihalcea, P. Tarau, E. Figa // Proceedings of The 20st International Conference on Computational Linguistics (COLING 2004), Switzerland, Geneva, August 2004.

[13] *Miller G.* Nouns in WordNet // WordNet – An Electronic Lexical Database / C. Fellbaum (ed.). The MIT Press. P. 23–47.

[14] *Resnik P.* Using information content to evaluate semantic similarity / P. Resnik // IJCAI 1995.

[15] *Snyder B., Palmer M.* The English all-words task / B. Snyder, M. Palmer // Proceedings of SENSEVAL-3. Third International workshop on the Evaluation of Systems for the Semantic Analysis of Texts. 2004. P. 41–43.

[16] *Vossen P., Rigau G., Alegria I., Agirre E., Farwell D., Fuentes M.* Meaningful results for Information Retrieval in the MEANING project / P. Vossen, G. Rigau, I. Alegria, E. Agirre, D. Farwell, M. Fuentes //Proceedings of Third International WordNet Conference, 2006.

Thesaurus-based Word Sense Disambiguation

Loukachevitch N., Chuiko D.

In the paper we describe a new method for word-sense disambiguation based on the Thesaurus of Russian Language RuThes. We evaluated precision of the algorithm for the «all-words» task and the task of thematic-oriented word-sense disambiguation.

For the «all-words» task the precision of our algorithm, which does not use sense-tagged corpora, is comparable with the results of the best systems of the specialized conference SENSEVAL-3. However the level of the precision for the «all-words» task is not enough for the use in information-retrieval applications.

For the task of thematic-oriented word-sense disambiguation the precision is much higher. Therefore it could be perspective to develop hybrid information-retrieval methods combining word-based techniques for all words and concept-based techniques for processing of thematic words and terms based on domain-specific thesauri or ontologies.