Устранение лексической многозначности терминов Википедии на основе скрытой модели Маркова

© Турдаков Денис

ВМК МГУ, ИСП РАН
turdakov@gmail.com

Аннотация

В статье описывается способ автоматического устранения лексической многозначности терминов естественного языка, использующий открытую энциклопедию Википедию. Рассматриваются проблемы применения существующих алгоритмов, и предлагается собственный метод, основанный на скрытой модели Маркова, параметры которой вычисляются на основе словаря и ссылочной структуры Википедии. Также, предлагается эвристика для ускорения описанного алгоритма, и приводятся экспериментальные оценки точности на различных тестовых корпусах.

1 Введение

Задача устранения лексической многозначности (word sense disambiguation) возникла в 50-х годах прошлого века, в качестве подзадачи машинного перевода. С тех пор, исследователи предложили огромное количество методов ее решения, однако она остается более чем актуальной и по сей день. В общем случае, задача включает в себя два аспекта:

1) фиксирование всех различных значений для каждого слова, относящегося к тексту;
2) определение способа выбора подходящего значения для каждого экземпляра слова.

Рассмотрим каждый из них подробнее. Существует два основных подхода к определению списка значений слов. Большинство работ опираются на предопределенные значения: списки слов, найденные в словарях, переводы на иностранные языки, и т. п. Вторым подходом является анализ способов употребления слов в различных источниках и выделение значений на основе этого анализа. Однако до сих пор ведутся споры о том, что является значением слова. Кроме того, часто необходимо определить значение не отдельно стоящего слова, а группы слов, образующих термин. Тогда задача осложняется еще и поиском терминов и определением, какие значения могут соответствовать каждому термину.

Алгоритмы для выбора подходящего значения используют два источника информации: контекст слова - информацию, которая содержится в тексте, в котором слово встретилось; и внешние источники, такие как словари и базы знаний. Современные методы можно разделить на два класса: методы, основанные на обучении по размеченным корпусам, и методы, основанные на внешних источниках знаний (тезаурусы, машинно-ориентированные словари, лексиконы). Хороший обзор алгоритмов можно найти в [1, 7], также краткий обзор алгоритмов, использующих Википедию, будет дан во второй части статьи.

Еще одной важной проблемой является оценка методов и их сравнение. Так как снятие многозначности обычно используется для улучшения работы большей системы, существует два способа оценки: in vitro - на сколько хорошо работают методы сами по себе - и in vivo - как снятие многозначности улучшает работу системы в целом. Для оценивания самих методов обычно используют два коэффициента: точность и полноту. **Точность** - это число слов, размеченных правильно, по отношению к числу слов, обработанных системой. **Полнота** - число слов, размеченных правильно, по отношению к числу слов в тестовом множестве. Также часто вводят **F-меру**, значением которой является среднее гармоническое между полнотой и точностью. Для сравнения методов снятия многозначности английских слов были разработаны тестовые наборы и проводятся конференции Senseval-1,2,3 и Semeval[20]. Эти тестовые наборы используют заранее определенные значения многозначных слов, которые берутся из словаря WordNet [15], это накладывает ограничение на возможность их использования. Так, методы использующие словарь Википедии, нельзя напрямую сравнить с методами, использующими словарь WordNet, так как количество значений слов в Википедии намного превосходит аналогичное число в WordNet. В работе [13] авторы смогли отобразить используемые значения на словарь WordNet, однако в дальнейшем [14] отказались от такой процедуры. Это связано с тем, что Википедия растет и изменяется очень

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

быстро, а ее словарь на порядок превосходит словарь WordNet.

1.1 Скрытая модель Маркова и задача устранения лексической многозначности

Проблема снятия лексической многозначности может быть переформулирована как задача максимизации с использованием формализма скрытых Марковских моделей.

Пусть T - множество терминов, M - множество значений, соответствующих терминам. Для последовательности терминов $\tau = t_1,...,t_n$, где $t_i \in \mathrm{T}, \forall i$, задача максимизации состоит в нахождении наиболее вероятной последовательности значений $\mu = m_1,...,m_n$, где $m_i \in \mathrm{M}, \forall i$, соответствующей входным терминам и согласованной с ограничениями модели:

$$\hat{\mu} = \arg\max_{\mu} P(\mu \mid \tau) = \\ = \arg\max_{\mu}\left(\frac{P(\mu)P(\tau \mid \mu)}{P(\tau)}\right) \quad (1)$$

Так как вероятность $P(\tau)$ постоянна для фиксированной входной последовательности, задача редуцируется к максимизации числителя равенства 1. Для решения этого уравнения делается Марковское предположение, что значение *i*-го термина зависит только от конечного числа значений предыдущих терминов:

$$\hat{\mu} = \arg\max_{\mu}\left(\prod_{i=1}^{n} P(m_i \mid m_{i-1},...,m_{i-k}) \cdot P(t_i \mid m_i)\right) \quad (2)$$

где k – порядок модели.

Множители уравнения 2 определяют скрытую Марковскую модель *k*-го порядка, где наблюдения соответствуют входным терминам, состояния соответствуют значениям терминов, $P(m_i \mid m_{i-1},...,m_{i-k})$ – модель перехода между состояниями и $P(t_i \mid m_i)$ – модель наблюдения, описывающая вероятность появления термина t_i в каждом состоянии m_i.

Несмотря на то, что рассматриваемую задачу нетрудно формализовать с помощью скрытой Марковской модели, дальнейшее использование этого формализма связано с проблемой разреженности языка. Так, чтобы построить модель перехода для Марковской модели первого порядка, необходимо оценить вероятность каждой пары состояний, что для задачи устранения лексической многозначности является вероятностью того, что два термина в конкретных значениях встретились вместе. Если для задачи определения частей речи слов, параметры Марковской модели можно оценить на основе сравнительно небольшого размеченного корпуса, то для задачи устранения лексической многозначности проблема оценки параметров сильно усложняется. Это связано с количеством значений. Так, Википедия содержит более двух миллионов концепций, кроме того задача осложняется тем, что частота употребления терминов в тексте распределена не по равномерному закону, а по закону Зипфа (Zipf law). Учитывая эти факты, несложно заметить, что для обучения Марковской модели потребуется размеченный корпус огромного размера. Ниже мы предложим способ оценки модели перехода для поставленной задачи с помощью семантической близости концепций Википедии, вычисленной на основе графа ссылок.

1.2 Википедия

Википедия - это открытая энциклопедия, создающаяся пользователями Веба. Сейчас англоязычная Википедия содержит более 2.5 млн. статей, не считая специальных страниц. Граф Википедии (вершины графа - это страницы энциклопедии, а ребра - гипертекстовые ссылки между ними) обладает малым диаметром и высоким коэффициентом кластеризации[12], что структурно отличает ее от тезауруса WordNet. Кроме того, Википедия содержит огромное количество дополнительной информации, которая используется для различных исследований. Рассмотрим структуру открытой энциклопедии более подробно.

Большинство статей, которые видят пользователи, имеют заголовок, обозначающий **концепцию**, описанную в теле статьи. В теле статьи пользователи ставят ссылки на концепции связанные с данной. Структурно ссылки состоят из двух основных частей: концепции энциклопедии, на которую указывает ссылка, и **термина**, который видит пользователь. Кроме обычных ссылок, бывают специальные, например, "Смотри также", которые указывают на концепции, сильно связанные с данной.

Кроме обычных статей, существует еще несколько типов страниц. Страницы устранения многозначности содержат списки значений многозначных слов. Эти страницы создаются, как и вся Википедия, вручную пользователями и поэтому содержат не все возможные варианты. При этом, среднее количество вариантов намного превосходит аналогичное количество в WordNet (табл. 1).

	Википедия Октябрь'08	**Википедия Март '09**	**WordNet**
Концепций	2 500 000	2 800 000	150 000
Многозначных терминов	260 000	290 000	26 000
Среднее кол-во значений	5,4	5,47	2,95

Таблица 1: Статистика Википедии и WordNet

Каждая статья Википедии принадлежит одной или нескольким категориям. Сами категории также

могут принадлежать более общим категориям. При этом они не образует таксономию, так как граф категорий может содержать циклы.

Еще одним важным типом статей являются редиректы. С этих страниц происходит автоматическое перенаправление на обычные статьи. По сути, эти статьи содержат синонимы концепций. Кроме того, Википедия содержит еще много дополнительных структур (шаблоны, инфобоксы, списки, и т.д.).

В следующем разделе будет дан обзор наиболее близких методов устранения многозначности. Далее будет описан разработанный метод и приведены оценки его работы.

2 Обзор существующих работ

Условно можно выделить три этапа развития методов устранения лексической многозначности. С 50-х по 80-е года были разработаны основные подходы, однако из-за отсутствия хороших машинных словарей и баз знаний, в этот период созданы только "игрушечные" системы, покрывающие только некоторые слова языка. Следующий этап, пик которого пришелся на 90-е годы, был обусловлен вручную созданными крупномасштабными базами знаний, такими как WordNet и CyC [4] и сбалансированными корпусами документов(Brown [5], Penn TreeBank [21]). Алгоритмы, разработанные в этот период, использовали структуру баз знаний или обучались на общепризнанных корпусах. Исследователи получили хорошие результаты, однако сложность ручного создания и поддержки в актуальном состоянии больших структур ограничила область применения этих алгоритмов. В начале 21 века исследователей в области обработки естественного языка заинтересовала возможность использования сетей документов, таких как WWW и Wikipedia, связанных гиперссылками, и созданных огромным числом независимых пользователей. Ниже приводится краткий обзор работ, в которых для решения задачи снятия лексической многозначности использовалась модель Маркова или Википедия.

Для оценки нижней границы точности методов устранения лексической многозначности обычно используют наивный метод, который всем словам присваивает их наиболее частое значение (baseline algorithm). Заметим, что этот метод эквивалентен Марковской модели нулевого порядка, в которой все термины равновероятны для любой фиксированной концепции.

Для методов, основанных на внешних источниках, в качестве нижней границы часто используют результат работы алгоритма Леска. Алгоритм Леска - это классический алгоритм автоматического снятия многозначности, введенный М. Леском в 1986 году [8]. Алгоритм основан на предположении, что многозначное слово и его окружение относятся к одной теме. Простая реализация алгоритма Леска состоит из трех шагов:

- Выбрать многозначные слова и их контексты
- Взять их определение в некотором словаре
- В качестве значения многозначного термина выбрать то, которое максимизирует количество общих слов в словарном определении данного значения и определений терминов контекста.

В работе [10] для обучения модели Маркова использовался корпус SemCor, и было показано, что его недостаточно, чтобы обучить даже модель первого порядка. Чтобы устранить проблему разреженности языка, авторы предложили использовать категории WordNet для построения модели перехода в Марковской цепи. Это позволило немного повысить точность алгоритма, в сравнении с наивным методом.

Аналогичный результат был получен и в более поздних работах [18, 19], где использовалась специализированная модель Маркова, с состояниями, хранящими части речи слов. Авторы этой работы показали, что без дополнительной семантической информации из внешних источников, корпуса SemCor недостаточно для нахождения параметров марковской модели для задачи устранения лексической многозначности слов, и наилучшие результаты показывает наивный метод.

В работах, использующих Википедию, продолжают просматриваться парадигмы, заложенные в 90-х годах прошлого века. Условно их можно охарактеризовать как методы, основанные на статистическом подходе, и методы, основанные на внешних данных.

В работе [13] Википедия использовалась как аннотированный корпус для обучения Наивного Байесовского классификатора. Многозначные термины и их значения выделялись из ссылок ([[статья-значение|многозначный термин]]). Для каждого многозначного термина, представленного в Википедии, строился вектор признаков, составленный из

- части речи многозначного термина
- локального контекста из трех слов слева и справа многозначного термина с их частями речи
- глобального контекста из пяти самых частых специфичных для значения слов, встречающихся во всевозможных контекстах термина.

Авторы вручную отобразили часть терминов Википедии в термины WordNet, чтобы провести эксперименты на общепринятых эталонных тестах корпуса SENSEVAL. Эксперименты показали, что Наивный Байесовский классификатор, обученный на Википедии, показывает лучшие результаты, чем алгоритм Леска и baseline algorithm. Кроме того было показано, что распределение терминов в

статьях Википедии сильно отличается от тренировочного множества корпуса SENSEVAL - вручную сбалансированного британского национального корпуса. Также эксперименты подтвердили предположение, что с ростом Википедии улучшается точность снятия многозначности.

Исследование, описанное в [13], получило развитие в системе автоматического обогащения документов ссылками на статьи Википедии [14]. В статье, посвященной этой системе, описываются методы для автоматического извлечения ключевых слов и снятия лексической многозначности на основе Википедии. Для выделения ключевых фраз использовался словарь, состоящий из заголовков Википедии, расширенный морфологическими формами, которые встречаются во внутренних ссылках Википедии на соответствующие статьи не менее 5 раз. Выделение ключевых фраз происходит в два этапа:

- поиск кандидатов,
- ранжирование ключевых фраз.

На первом этапе выделяются всевозможные N-граммы, присутствующие в словаре, после этого на втором этапе выделенным терминам присваиваются веса на основе одного из трех методов: tf-idf, хи-квадрат, и информативность (вероятность использования термина в качестве текста ссылки).

На основании результатов проведенных экспериментов утверждается, что последний метод показывает наилучшие результаты. Во второй части работы приводится алгоритм снятия многозначности в выделенных ключевых фразах. Авторы использовали комбинацию алгоритма Леска [8] и статистического алгоритма, обученного на Википедии. В алгоритме Леска в качестве словарного определения термина бралась соответствующая статья Википедии, а в качестве контекста абзац, в котором встретился термин. В качестве статистического алгоритма использовался классификатор[13], описанный выше. Наконец, предполагая ортогональность этих методов, авторы использовали расхождения в результатах, как признак потенциальной ошибки и игнорировали такой результат. Оценка получившейся системы производилась на тестовом наборе из 85 случайно выбранных статей Википедии, специально размеченных вручную. Эксперименты показали, что статистический метод показывает большую точность (92.91%) и полноту (83.1%), чем алгоритм Леска (80.1% и 71.86% соответственно), а комбинирование алгоритмов увеличивает точность (94.33), но снижает полноту (70.51%).

Работа [3] является развитием алгоритма Леска, с учетом дополнительной информации, которую можно извлечь из Википедии. Для создания словаря и поиска возможных значений терминов использовались заголовки статей, редиректы, страницы значений многозначных слов и текст ссылок. Для каждого термина словаря собирался вектор признаков, состоящий из тэга категории, контекста (слова или термина встречающегося вместе с данным термином) и класса термина (Человек, Место, Организация, Остальное). В качестве значения многозначного термина выбирался кандидат, максимально похожий на контекст, где похожесть вычислялась как косинус между векторами признаков.

В работе [2] также рассматривается использование векторной модели для снятия многозначности имен собственных, однако, в отличии от [3], практически не уделяется внимания нахождению различных признаков, а, в дополнение к векторной модели, предлагается использовать иерархию категорий Википедии для обучения линейного классификатора, основанного на методе опорных векторов (Support Vector Machine, SVM).

В работе [11] для снятия лексической многозначности используется семантическая близость терминов [16]. Расстояние между терминами вычисляется на основе графа ссылок Википедии. Для каждого возможного значения термина вычисляется близость до терминов контекста и выбирается наиболее близкое. Авторы показали, что их алгоритм работает лучше, чем алгоритм Леска. Для оценки качества алгоритма использовалось тестовое множество, созданное на основе аннотаций ссылок из статей Википедии.

Для снятия многозначности авторы [17] использовали подход, основанный на машинном обучении в комбинации с семантической похожестью, описанной в [16]. В качестве тренировочного множества использовалось 500 случайных статей Википедии. Положительными примерами устранения многозначности служили термины, на которые указывали ссылки, а остальные возможные значения, как и в [11] полученные из ссылок Википедии, служили отрицательными примерами. В качестве признаков использовались вероятность значения многозначного термина, полученная просмотром всех ссылок Википедии, и расстояние до терминов контекста. Для вычисления расстояния до контекста использовалась та же мера, что и в [11]. Но, кроме того, терминам контекста придавался вес, посчитанный как среднее между информативностью данного термина и близостью термина к центральной нити документа, определенной как средняя близость между текущим термином и остальными терминами контекста. Еще одним признаком послужило качество контекста, определенное как сумма весов терминов, посчитанных на предыдущем шаге. Основываясь на данных признаках, авторы провели сравнительное тестирование нескольких алгоритмов машинного обучения (Naive Bayes, SVM, C4.5 с вариациями) на тестовом множестве из 100 случайных статей Википедии, и показали, что данный подход дает лучшие результаты (97.1%), чем описанный в [11].

В работе [22] так же, как и в [11] и [17] используется мера семантической близости терминов. Однако, для ее вычисления авторы

предложили использовать информацию о типах ссылок и давать им разный вес. Различные меры близости на взвешенном графе сравниваются между собой на примере задачи снятия лексической многозначности, и показывается, что коэффициенты Дайса и Жаккара дают наилучшие результаты. Мы используем аналогичный метод вычисления близости терминов для оценки параметров модели.

3 Снятие многозначности

Методы, предложенные в вышеперечисленных работах, основанные на внешних знаниях, имеют один общий недостаток. Они неявно используют предположение, что в тексте существуют однозначные термины, на основании значений которых впоследствии определяются значения многозначных терминов. Однако, было замечено, что с ростом Википедии словарь многозначных терминов увеличивается, причем дополнительные значения появляются у наиболее употребляемых терминов. Это приводит к тому, что в неспецифических сообщениях, таких как новостные статьи, все термины имеют более одного значения, либо встречающиеся однозначные термины мало связаны с основной темой документа. Это ухудшает точность алгоритмов, основанных на однозначном контексте, и ведет к необходимости разработки метода, лишенного такого недостатка.

Далее будет приведено описание такого метода, основанного на скрытой модели Маркова. Вероятности модели перехода мы оцениваем с помощью семантической близости концепций, способ подсчета которой описан в следующем подразделе. Вероятности модели наблюдений и априорная вероятность значений оценивается с помощью эмпирического распределения значений и терминов во внутренних ссылках Википедии. Кроме того ссылки, совместно со словарем Википедии и специальными статьями используются для создания словарей терминов и их значений (раздел 3.2). Далее приводится алгоритм и результаты экспериментов на различных корпусах.

3.1 Семантическая близость концепций

Ранее [22], мы разработали простую меру близости между статьями Википедии, которая может быть полезна для различных задач, в том числе для снятия многозначности. Меры близости между вершинами графа, описанные в литературе, можно разделить на два широких класса: меры, основанные на локальной информации, такие как косинус угла между векторами, коэффициенты Дайса и Жаккара, соцетирование и т. п., и меры, основанные на распространяющейся активации (spreading activation), например, SimRank [6]. В то время как меры, относящиеся ко второму классу, показывают более качественные результаты, их вычислительная сложность ($O(n^3)$ для SimRank [9]) не позволяет использовать их для работы с большими объемами данных. Кроме того, вычисление семантической близости между двумя вершинами требует построения полной матрицы близости для всех вершин графа. Поэтому в нашей работе мы используем меру, основанную на коэффициенте Дайса, часто используемую в системах информационного поиска.

Для двух статей Википедии мера Дайса определяется как удвоенное отношение числа их общих соседей к общему числу всех соседей обоих статей. Формально

$$sim(A,B) = Dice(A,B) = \frac{2\times|n(A)\cap n(B)|}{|n(A)|+|n(B)|}, \quad \textbf{(3)}$$

где $n(X)$ – множество статей, связанных ссылкой со статьей *X*.

Мы исследовали структуру Википедии и заметили, что некоторые типы ссылок чрезвычайно релевантны по отношению к семантической близости, в то время как другие могут привести к неверным результатам. Поэтому, в дополнение к основной мере, мы ввели схему весов, основанную на типах ссылок. Подробное описание способа вычисления семантической близости здесь не приводится, так как не относится к основной теме статьи, и его можно найти в работе [22].

3.2 Создание словарей

На данный момент Википедия содержит статьи, описывающие более 2.5 миллионов концепций. Мы используем названия статей для создания словарей, которые используется для поиска терминов в текстах. После того, как все термины в тексте найдены, они представляются как последовательность наблюдений, а их значения - как соответствующие состояния в скрытой модели Маркова.

Для формирования словаря терминов мы берем названия всех статей, описывающих соответствующие концепции и названия всех страниц переадресации на эти статьи.

Основным источником значений терминов является категория "Disambiguation pages". Статьи, входящие в эту категорию, содержат списки возможных значений и ссылки на страницы, описывающие эти значения. Однако в Википедии не существует четких правил для создания таких страниц, поэтому часто они содержат много лишних ссылок, напрямую не связанных с многозначной концепцией. Поэтому при обработке этих страниц мы выделяем только те значения, которые содержат в своем названии словоформы многозначной концепции, или для которых она является акронимом. Эта эвристика очень жесткая и отсеивает много хороших значений (мы добавляем их позднее при анализе ссылок). Например, для термина "*NATO*" будет найдено значение "***N**orth **A**tlantic **T**reaty **O**rganisation*" но пропущено значение "*Mora (plant)*" – растение, которое часто называют "нато".

Система заранее не имеет информации, какие тексты придется анализировать, следовательно, не должна быть чувствительна к регистру слов. Поэтому все слова в словаре мы приводим к верхнему регистру. Википедия, напротив, чувствительна к регистру, и одинаковые термины, написанные в разном регистре, могут указывать на различные концепции. Для решения этой проблемы, мы добавляем такие концепции в словарь значений терминов и выбираем нужной значение на этапе устранения многозначности терминов текста.

Кроме того, большое количество терминов содержит в названии уточняющие концепции, например *"Platform (computing)"*. Мы убираем текст в скобках и в случае коллизий применяем подход, аналогичный тому, который используется при приведении к одному регистру.

Во введении упоминалось, что источником значений многозначных терминов могут быть как словари, так и способы употребления терминов в текстах. В нашем случае дополнительным источником значений являются ссылки между страницами. Любая ссылка содержит две части: текст, который видит пользователь, и концепцию Википедии, на которую в действительности ведет ссылка. Мы анализируем употребление всех терминов из созданного словаря в качестве текста ссылок и добавляем в список значений этих терминов концепции, на которые указывали данные ссылки. На этом заканчивается формирование словаря значений терминов.

В итоге словарь терминов содержит более 5 500 000 терминов, соответствующих 2 500 000 концепций, из них многозначных терминов – 260 000 элементов, а среднее количество значений многозначных терминов равно 5,4.

3.3 Описание алгоритма

Во введении было показано, как использовать формализм скрытых моделей Маркова для решения задачи устранения лексической многозначности. Основная сложность использования этого формализма состоит в оценке параметров модели. Воспользуемся информацией Википедии, чтобы решить эту проблему.

Для оценки модели наблюдения воспользуемся ссылками Википедии. На основании способа построения словарей можно заметить, что термины, соответствующие синонимам концепции, могут появиться только из заголовка статьи, описывающей концепцию, названий редиректов на концепцию и терминов, совпадающих с текстом ссылок на концепцию. Исходя из этого, определим условную вероятность термина t_i^j, соответствующего значению m_i через эмпирическую вероятность $\hat{P}(t_i^j \mid m_i)$:

$$P(t_i^j \mid m_i) = \hat{P}(t_i^j \mid m_i) = \frac{C(t_i^j, m_i)}{C(m_i)} \quad \textbf{(4)}$$

где $C(t_i^j, m_i)$ – количество ссылок на концепцию m_i, в которых термин которых совпадал с t_i^j, включая редиректы и название концепции, как специальный тип ссылок, а $C(m_i)$ – общее количество ссылок на концепцию.

Чтобы оценить модель перехода сделаем предположение, что вероятность значения m_i, при условии предыдущего контекста $m_{i-1}...m_{i-k}$ пропорциональна линейной комбинации близости значения к контексту и априорной вероятности этого значения.

$$P(m_i \mid m_{i-1}..m_{i-k}) = \hat{P}(m_i \mid m_{i-1}..m_{i-k}) = \alpha\left(sim(m_i; m_{i-1}..m_{i-k}) + \beta * P(m_i)\right) \quad \textbf{(5)}$$

Для модели первого порядка близость значения к контексту, соответствующему предыдущему значению, вычисляется через коэффициент Дайса, способом, описанным в разделе 3.1. Чтобы оценить близость значения к контексту из нескольких терминов, представим их в виде обобщенной концепции, объединяющей все входящие в нее значения, тогда

$$n(B_1 B_2 ... B_m) = \bigcup_{i=1}^{m} n(B_i), \quad \textbf{(6)}$$

и близость вычисляется так же, как и для двух обычных концепций.

Априорную вероятность значения будем оценивать на основе ссылок, способом аналогичным тому, который мы использовали при оценке модели наблюдения.

$$P(m_i) = \hat{P}(m_i) = \frac{C(m_i)}{\sum_i C(m_i)} \quad \textbf{(7)}$$

Коэффициент нормализации α в уравнении 5 не влияет на решение задачи максимизации, поэтому его можно не учитывать. Коэффициент β на основании экспериментов мы определили равным 1.

После определения всех параметров модели задача максимизации решается с помощью алгоритма Витерби. Этот алгоритм использует замечание, что наиболее вероятный путь до каждого следующего состояния зависит только от наиболее вероятного пути через k предыдущих состояний. Таким образом, количество сравнений на каждом шаге экспоненциально зависит от k и равно

$$\prod_{i=k-n}^{k-1} | m_i |$$

Чтобы сократить время работы алгоритма, мы выдвинули наивное предположение, что **наиболее вероятный путь до состояния m_i зависит от**

k **последних терминов наиболее вероятного пути до состояния** $\mathbf{m}_{i-1}$**.** В этом случае каждое состояние должно хранить дополнительную информацию не более чем о k предыдущих терминах, и, таким образом, модель сведется к специализированной Марковской модели первого порядка. Наиболее вероятная последовательность состояний для такой модели находится так же, как и для обычной Марковской модели первого порядка, за исключением вычисления вероятности перехода между состояниями.

Конечно, это предположение в общем случае неверно, однако в рамках данной задачи, оно позволяет уменьшить порядок модели и, при этом, учесть контекст из нескольких терминов, тем самым не сильно ухудшить точность метода (табл. 3 и 4).

4 Эксперименты

4.1 Коллекция для тестирования

Как уже упоминалось, для оценки точности и полноты методов обычно используются тестовые коллекции Senseval-1,2,3 и SemEval, основанные на WordNet. Однако различия между Википедией и WordNet не позволяют использовать их напрямую. Более того, отображение концепций Википедии на словарь WordNet (что само по себе является трудоемкой задачей, также требующей оценки) не дает возможности корректно сравнить алгоритмы из-за взаимной неоднозначности такого отображения. Поэтому в методах, основанных на Википедии, в качестве тестового корпуса часто используются сами статьи Википедии, причем обрабатываются только термины, представленные в виде ссылок, а значениями этих терминов служат концепции Википедии, на которые указывают ссылки. Несложно заметить, что для таких тестовых корпусов методы, основанные на машинном обучении и обученные на Википедии, дают наилучшие результаты из-за схожести распределений обучающего и тестового множеств [17].

Для оценки нашего метода мы создали тестовое множество, выделив 500 случайных статей Википедии. Использовались только статьи, описывающие однозначные концепции, так как они наиболее близки к неструктурированным текстам. Кроме этого, для составления более полной картины, мы вручную разметили тестовую коллекцию из 131 документа, состоящую из новостных сообщений, взятых из различных источников, и нескольких научных статей. Характеристики коллекций представлены в таблице 2.

Среднее число значений многозначных терминов в обеих коллекциях сильно превышает аналогичное число во всем языке (табл. 1). Это происходит потому, что у часто употребляемых терминов значений больше. Кроме того, процент многозначных терминов в коллекции, размеченной вручную, намного превышает аналогичное число в коллекции, автоматически созданной из статей Википедии.

	Новости и научные статьи	Статьи Википедии
Количество документов	131	500
Количество терминов	8236	50974
Многозначных терминов	6952	39332
Среднее количество значений	22,34	35,34

Таблица 2: Характеристики тестовых коллекций

Также следует заметить, что с изменением Википедии приходится изменять и тесты, так как появляются новые термины, и увеличивается количество значений. И если создать тестовую коллекцию по Википедии можно автоматически, то тесты, созданные вручную, придется заново вручную переразметить.

4.2 Результаты

Результаты экспериментов представлены в таблицах 3 и 4. Все эксперименты проводились на снимке Википедии, полученном в октябре 2008г. Алгоритм применялся ко всем найденным терминам текста, поэтому точность и полнота совпадают.

Порядок	Модель Маркова	MM с эвристикой
0	53.12	53.12
1	54.00	54.00
2	54.50	54.49
3	54,76	54.72

Таблица 3: Результаты работы алгоритма на коллекции новостей и научных статей

Порядок	Модель Маркова	MM с эвристикой
0	91,34	91,34
1	91,64	91,64
2	92,40	92.37
3	92,51	92,41

Таблица 4: Результаты работы алгоритма на коллекции статей Википедии

Сравнительно низкие результаты, полученные на первом корпусе, связаны с тем, что мы считали заведомо неверным ответ алгоритма, данный для терминов, не имеющих правильного значения среди концепций Википедии. Такими терминами, в основном, являются имена людей и слова, входящие в устойчивые выражения, например, слово “lot” в выражении “a lot of time…”. Если не учитывать такие термины, точность алгоритма достигает 76,84%.

Для сравнения алгоритм, описанный в работе [22] показывает точность 43,41% и полноту 34,77% на первом тестовом наборе и, соответственно, 79,58% и 77,29% на наборе из статей Википедии. На снимке, сделанном в июле 2008г., этот алгоритм давал точность и полноту 59,19% и 49,60% на первой коллекции и 91,93% и 89,62% на второй. Эти результаты наглядно демонстрируют, как с ростом Википедии ухудшается точность алгоритмов, использующих однозначный контекст.

Наилучшие результаты были представлены в работах [14] и [17] (94,33/70,51 и 98.4/95,7) и получены на аналогичных коллекциях, основанных на статьях Википедии. Однако, эти алгоритмы так же используют однозначный контекст, что, несомненно, ухудшит их точность и полноту при использовании с новыми версиями Википедии.

5 Заключение

В работе предлагается метод устранения лексической многозначности терминов естественного языка, основанный на Марковской модели, параметры который вычислены с помощью данных Википедии. Проблема разреженности языка решается предположением, что апостериорная вероятность значения термина при условии предыдущего контекста пропорциональна линейной комбинации семантической близости соответствующих концепций Википедии и априорной вероятности значения. Для ускорения алгоритма предложена эвристика, которая, в рамках поставленной задачи, дает выигрыш по времени выполнения, при этом незначительно ухудшая точность результата.

Однако, анализ ошибок алгоритма позволил выявить существенный недостаток: метод неявно предполагает, что все термины в тексте имеют общий смысл. Однако, часто в тексте кроме основной семантической линии существует несколько параллельных, таких как место и время основных событий. Основываясь на этом замечании, мы пришли к выводу, что применение данного алгоритма необходимо комбинировать с методом, выделяющим семантически связанные цепочки терминов (lexical chains). В этом направлении мы планируем сделать следующий шаг работы.

Литература

[1] Eneko Agirre, Philip Glenny Edmonds. Word Sense Disambiguation: Algorithms and Applications. Springer, 2006

[2] Razvan Bunescu, Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Trento, Italy, April 2006

[3] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In Proc. 2007 Joint Conference on EMNLP and CNLL, pages 708–716, Prague, The Czech Republic, 2007.

[4] Cycorp, Inc. Web site. www.cyc.com

[5] Francis, W. and Kucera, H. Brown Corpus Manual. http://icame.uib.no/brown/bcm.html

[6] Glen Jeh, Jennifer Widom, SimRank: a measure of structural-context similarity, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002

[7] Nancy Ide and Jean Vйronis. Word sense disambiguation: The state of the art. Computational Linguistics, 1998

[8] Michael Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, ACM Special Interest Group for Design of Communication Proceedings of the 5th annual international conference on Systems documentation, p. 24 – 26, 1986.

[9] D. Lizorkin, P. Velikhov, M. Grinev and D. Turdakov. Accuracy Estimate and Optimization Techniques for SimRank Computation. In VLDB '08: Proceedings of the 34th International Conference on Very Large Data Bases, pages 422--433.

[10] C. Loupy, M. El-Beze, and P. F. Marteau. 1998.Word Sense Disambiguation using HMM Tagger. In *Proceedings of the 1st International Conference* on Language Resources and Evaluation, LREC, pages 1255–1258, Granada, Spain, May.

[11] Olena Medelyan, Ian. H. Witten and David Milne. Topic Indexing with Wikipedia. Proc. AAAI'08 Workshop on Wikipedia and Artificial Intelligence

[12] Menczer Filippo. Evolution of document networks. Proceedings of the National Academy of Sciences of the United States of America.

[13] Rada Mihalcea. Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT 2007, pages 196–203, Rochester, NY, April 2007

[14] Mihalcea, R. and Csomai, A. (2007) Wikify!: linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge management (CIKM'07)

[15] George A. Miller, WordNet: a lexical database for English, Communications of the ACM, v.38 n.11, p.39-41, Nov. 1995

[16] David Milne, Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proc. AAAI'08 Workshop on Wikipedia and Artificial Intelligence

[17] David Milne, Ian H. Witten. Learning to Link with Wikipedia. Proceedings of the ACM Conference on Information and Knowledge Management, 2008
[18] Antonio Molina and Ferran Pla and Encarna Segarra and Lidia Moreno. Word Sense Disambiguation using Statistical Models and WordNet. Proceedings of 3rd International Conference on Language Resources and Evaluation, LREC2002, Las Palmas de Gran Canaria
[19] Molina, F. Pla, E. Segarra, WSD system based on specialized Hidden Markov Model (upv-shmm- aw), in: R. Mihalcea, P. Edmonds (Eds.), Senseval: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 171-174
[20] Senseval. Web site. www.senseval.org
[21] The Penn Treebank Project. http://www.cis.upenn.edu/~treebank/
[22] D. Turdakov, P. Velikhov. Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In proceedings of SYRCoDIS, 2008.

Sense disambiguation of Wikipedia terms based on Hidden Markov Model

Denis Turdakov

The paper presents a method for word sense disambiguation using external knowledge extracted from the open encyclopedia Wikipedia. We analyse the drawbacks of the existing word sense disambiguation algorithms and propose own algorithm, based on Hidden Markov Model, to overcome these drawbacks. HMM parameters are estimated by empirical probabilities derived from the Wikipedia dictionary and link structure. A heuristics for speeding up the computational aspects of the algorithm is proposed, and the evaluation of the algorithm for several test collections is provided.