ВЕРОЯТНОСТНЫЙ ПОДХОД К ЗАДАЧЕ РАЗРЕШЕНИЯ ОМОНИМИИ СЛОВ И СЛОВАРНЫХ ПАР

A PROBABILISTIC APPROACH TO LEXICAL AMBIGUITY RESOLUTION OF WORDS AND WORD PAIRS

Баглей С.Г. *(baglei@galaktika.ru),* ***Антонов А.В****. (alexa@galaktika.ru),*
Мешков В.С. *(meshkov@galaktika.ru),* ***Титов А.В****. (titov@galaktika.ru)*
Корпорация "Галактика", г. Москва

В статье описан метод снятия частичной омонимии слов и приведения к правильной основной форме словосочетаний, образующих ИнфоПортрет выборки документов в системе "Галактика-Zoom". Метод основан на анализе частотности употребления "проблемных" языковых инвариантов и использует статистические данные текстовых массивов и их элементов.

1. Введение

Одной из важных особенностей информационно-аналитической системы "Галактика-Zoom", является возможность формирования Информационного Портрета (ИнфоПортрета) выборки документов. ИнфоПортрет представляет собой список языковых инвариантов (слов и словосочетаний), отличающих данную выборку от прочих. Технология построения Информационного Портрета подробно описана в работах [1, 2, 3]. Основой для работы данной технологии являются статистические и лингвистические методы обработки текстовой информации.

Далее приведен пример ИнфоПортрета выборки документов по запросу *"ЗАПАДНАЯ ЦИВИЛИЗАЦИЯ"*, проведенного на массиве документов, представляющих собой публикации русскоязычных СМИ за период 2004-2007 годов:

ЗАПАДНАЯ ЦИВИЛИЗАЦИЯ	РУССКИЙ	ИМПЕРИЯ
ЦИВИЛИЗАЦИЯ	ЛИБЕРАЛЬНЫЙ	ГЛОБАЛИЗАЦИЯ
ЗАПАДНЫЙ	СВОБОДА	ИДЕОЛОГИЧЕСКИЙ
ЗАПАД	ТЕРРОРИЗМ	ГЛОБАЛЬНЫЙ
ИСЛАМСКИЙ	ЕВРОПЕЙСКИЙ	БЛИЖНИЙ ВОСТОК
ИСЛАМ	ХОЛОДНАЯ ВОЙНА	ХРИСТИАНСКИЙ
МУСУЛЬМАНИН	ЗАПАДНЫЙ МИР	ИСТОРИЧЕСКИЙ
ЦЕННОСТЬ	ЭЛИТА	СОЗНАНИЕ
МУСУЛЬМАНСКИЙ	НАЦИЯ	ДУХОВНЫЙ
ДЕМОКРАТИЯ	АМЕРИКАНСКИЙ	ГЕОПОЛИТИЧЕСКИЙ
РЕЛИГИОЗНЫЙ	СОЕДИНЕННЫЕ ШТАТЫ	РЕВОЛЮЦИЯ
РЕЛИГИЯ	ДЕМОКРАТИЧЕСКИЙ	ЗАПАДНОЕ ОБЩЕСТВО
ИДЕОЛОГИЯ	ЧЕЛОВЕЧЕСТВО	СССР
ИСЛАМСКИЙ МИР	МИРОВОЙ	РУССКАЯ ЦИВИЛИЗАЦИЯ
ЦИВИЛИЗАЦИОННЫЙ	ВОСТОК	ПОЛИТИКА
АМЕРИКА	МИРОВАЯ ВОЙНА	ИРАК

Таблица 1. ИнфоПортрет выборки документов по запросу "ЗАПАДНАЯ ЦИВИЛИЗАЦИЯ"

2. Описание проблемы и постановка задачи

При формировании ИнфоПортрета в нем часто встречались отдельные его элементы - слова, резко отличающиеся от тематики, к которой относился весь ИнфоПортрет. Например, по запросу "*МУЗЫКА И КОНЦЕРТ*" один из фрагментов сформированного ИнфоПортрета представлялся в таком виде:

МАДОННА
ТАНЦЕВАЛЬНЫЙ
ПЕРВЫЙ АЛЬБОМ
ПОПУЛЯРНЫЙ ИСПОЛНИТЕЛЬ
ЭТА ПЛОЩАДЬ
СЧАСТЬЕ
КАЖДАЯ ПЕСНЯ
ЕГО КОНЦЕРТ
АТТРАКЦИОН
ЛЕТНИЙ
САНТА
САКСОФОН
МЕНЬ
КАЗАЧЕНКО

Таблица 2. Омонимия в ИнфоПортрете по запросу "МУЗЫКА И КОНЦЕРТ"

Очевидно, что элемент ИнфоПортрета "*МЕНЬ*" резко выделяется на фоне общей тематики ИнфоПортрета. Можно сделать предположение, что в проанализированном тексте существует словоформа "*МЕНЯ*", относящаяся как к парадигме имени собственного "*МЕНЬ*" (в родительном падеже), так и к парадигме местоимения "*Я*" (в родительном падеже).

В качестве другого примера появления в ИнфоПортрете резко отличающегося элемента можно привести ИнфоПортрет, сформированный по запросу "*ПОПЕЧИТЕЛЬСТВО ИЛИ УСЫНОВЛЕНИЕ*":

РЕБЕНОК
РОДИТЕЛЬ
УСЫНОВЛЕНИЕ
ПОПЕЧИТЕЛЬСТВО
СОВМЕСТНАЯ СОБСТВЕННОСТЬ
ДИТЯ
ПОПЕЧЕНИЕ
ОСТАТЬСЯ
УСЫНОВИТЕЛЬ
ВОСПИТАНИЕ
ОПЕК
УСЫНОВИТЬ
ГРАЖДАНСКОЕ СОСТОЯНИЕ
АНАЛОГИЧНЫЕ УЧРЕЖДЕНИЯ

Таблица 3. Омонимия в ИнфоПортрете по запросу "Попечительство или усыновление"

Можно заметить, что элемент ИнфоПортрета "*ОПЕК*" в данном контексте неуместен. Вероятно, такой элемент появился по причине того, что словоформа "*ОПЕКА*" входит в парадигму аббревиатуры – имени собственного "*ОПЕК*" (в родительном падеже). Вместе с тем, данная словоформа принадлежит парадигме существительного "*ОПЕКА*" (в именительном падеже). А далее на этапе приведения слова к основной форме выбиралась неверная парадигма. То есть, очевидно, что проблема в работе модуля морфологического разбора при формировании ИнфоПортрета возникала из-за наличия в тексте документов неполных (частичных) омонимов - слов одной и той же части речи, у которых совпадает не вся система словоформ.

При работе со словосочетаниями (парами слов) в дополнение к омонимичности возникала проблема несколько иного рода. Исходя из сложившейся традиции словоупотребления, для некоторых словосочетаний уже сформировались устойчивые языковые конструкции. И для таких конструкций употребление единственного или множественного числа одного из слов, их составляющих, является важным элементом, часто определяющим смысл всей конструкции. Изменение данного элемента при приведении словосочетания к основной форме может быть грамматически правильным, но не употребляться или употребляться в ином значении. Например, *ВООРУЖЕННЫЕ СИЛЫ*, *МИРОТВОРЧЕСКИЕ ВОЙСКА*, *ВЗРЫВНЫЕ РАБОТЫ* – словосочетания, имеющие определенный смысл при их употреблении только во множественном числе. Основная форма словосочетаний *ВООРУЖЕННАЯ СИЛА*, *МИРОТВОРЧЕСКОЕ ВОЙСКО* в единственном числе может употребляться, но при

этом смысл словосочетаний будет искажен. То есть, можно допустить, что для перечисленных словосочетаний подобное приведение к основной форме в ИнфоПортрете будет ошибочным.

Выбор неверной основной формы словосочетаний при формировании ИнфоПортрета возникает в нескольких случаях, каждый из которых специфичен для русского или английского языков, с которыми работает система “Галактика-Zoom”. В первом случае, проявляющемся в текстах на русском языке, словосочетание может употребляться только во множественном числе. Так как существительное, входящее в словосочетание, само по себе может употребляться в единственном числе, то и все словосочетание может быть приведено к соответственной основной форме.

Другим случаем неверного приведения к основной форме является ситуация, характерная для английского языка. Распространенным явлением в английском языке является совпадение множественного числа существительного с третьей формой глагола. Например, в словосочетании *CLINTON LEAVES - КЛИНТОН ПОКИДАЕТ*, третья форма глагола *TO LEAVE* совпадает по написанию с множественным числом существительного *A LEAF*. Результатом приведения словосочетания *CLINTON LEAVES* к основной форме был элемент ИнфоПортрета *CLINTON LEAF.* Дословный перевод полученного результата – *ЛИСТ КЛИНТОНА* – полностью искажает смысл словосочетания. Другим примером проявления описываемой проблемы являлось приведение словосочетания *BARGE SINKS – БАРЖА ТОНЕТ* к основной форме – *BARGE SINK*, то есть, в дословном переводе, *СЛИВНОЙ СТОК НА БАРЖЕ*. В данном случае смысл словосочетания также теряется.

Таким образом, можно выделить две основные проблемы, связанные с неоднозначностью в текстах: проблема омонимии слов и проблема приведения словосочетаний к неверной основной форме. Далее описан способ решения перечисленных проблем, который был реализован в системе “Галактика-Zoom”.

3. Метод решения проблем омонимии и приведения к неверной основной форме

3.1. Общие требования к методу

Несмотря на то, что уже существуют и доказали свою эффективность алгоритмы снятия омонимии, основанные на анализе контекста омонима или построении семантических сетей, моделирующих структуру текста, и реализованных в различных системах текстового анализа, они оказались сложно реализуемыми в имеющихся условиях работы системы.

При решении данной задачи одним из требований являлось сохранение быстродействия поисковой системы. Объемы реальных текстовых массивов, которые требуется обработать и проанализировать в системе “Галактика-Zoom”, достаточно большие – в некоторых случаях они достигают 9 миллионов документов. Учитывая этот факт, операции обработки и анализа необходимо выполнять достаточно быстро, чтобы не вынуждать пользователя системы специально ожидать выдачи результатов поискового запроса. Лингвистические методы, основанные на разборе текста или ближайшего окружения (контекста), содержащего омонимичные слова или словосочетания, требуют существенных временных затрат и выделения достаточно больших вычислительных ресурсов.

Вместе с тем, нам были доступны статистические данные о словах и словосочетаниях, формируемые после проведения этапа предварительной обработки документов. Мы решили использовать для решения возникшей проблемы омонимии.

Задачи снятия частичной омонимии и выбора правильной основной формы различны по своей природе и требуют различных путей для их решения. При этом, стоит учесть, что в вышеперечисленных задачах есть и общее. Далее вместе с общими ограничениями описаны специфические особенности, присущие каждой из задач.

3.2. Снятие частичной омонимии слов

При выборе модели для решения описанной проблемы мы исходили из следующего свойства языка. Омонимичные слова достаточно редко употребляются в одном контексте. Природа данного свойства не является предметом исследования, оно лишь было основой для развития метода.

Исходя из этого, существует статистическая зависимость совместного употребления омонимичных слов. Эту зависимость можно наблюдать и основываться на ней при решении задачи. Для словосочетаний вероятность совместного употребления омонимов в одном контексте еще менее вероятна, так как встречаемость устойчивых словосочетаний реже, чем у отдельных слов. ИнфоПортрет также формируется на основе документов, составляющих контекст запроса. Следовательно, статистика слов и словосочетаний, составляющих ИнфоПортрет, сохраняет описанное свойство языка, и мы можем применить частотный анализ статистики для выявления зависимостей.

В целом, метод основан на выборе наиболее частотной словоформы из выборки текстов. При обнаружении в тексте словоформы, входящей в “проблемный” набор, мы рассчитываем наиболее частотную словоформу данного слова в выборке текстов. Найденная словоформа с наибольшей частотностью будет считаться наиболее

типичной для данной выборки; будем принимать ее в качестве "рабочей". Найденная "рабочая" словоформа приводится к первой форме для включения в ИнфоПортрет. Далее опишем модель нахождения словоформы.

Обозначим через *w* словоформу, совпадающую с различными формами из словоизменительных парадигм нескольких слов. Пусть, *V* – набор основных форм таких слов, в парадигме которых существует омонимичная словоформа *w*. Тогда, из набора *V* получаем набор *F* пересекающихся (общих) словоформ в наборе парадигм слов, основными формами которых являются словоформы, составляющие набор *V*. Частоты основных форм из набора *V* в базе документов обозначим:

Воспользуемся теоремой гипотез (формулой Байеса) для решения задачи. В нашем случае, полная группа несовместных гипотез $H_{w,1}, H_{w,2}, \ldots, H_{w,|V|}$ представляет собой гипотезы, при которых омонимичная словоформа *w* может быть отнесена к одной из словоформ из набора *V*.

$H_{w,i} \in H_{w,1}, H_{w,2}, \ldots, H_{w,|V|}$, – гипотеза отнесения словоформы к некоторой основной форме из набора *V*, где $i = \overline{1,|V|}$.

Априорные вероятности этих гипотез обозначим соответственно:

$$P(H_1), P(H_2), \ldots, P(H_{|V|}) \qquad (1).$$

Принимаем априорные значения вероятностей равными, то есть:

$$\begin{cases} P(H_{i,w}) = \dfrac{1}{|V|}, \\ i = \overline{1,|V|}. \end{cases}$$

При первом цикле анализа документов определяем частоты встречаемости в базе словоформ, соответствующих гипотезам из набора (1). Принимаем полученные частоты в качестве наступления события. То есть, обозначим через q_W частоту некоторой омонимичной словоформы *w* в базе документов;
q_i - частота *i*-ой основной формы, $i = \overline{1,|V|}$.

Считаем наступлением события найденные частоты для основных форм и омонимичных словоформ.

Далее, на втором цикле анализа документов уточняем априорные значения вероятностей. Находим условную вероятность отнесения словоформы *k*+1 из набора *F* к основной форме *i* из набора *V* следующим образом:

$$\begin{cases} P(H_{i,k+1}) = P(H_{i,k}) * \dfrac{q_i}{P(H_{i,k}) * q_w + q_i}, \\ i = \overline{1,|V|}, \\ k = \overline{1,|F-1|} \end{cases} \qquad (2).$$

После уточнения вероятностей гипотез выбираем гипотезу с максимальной вероятностью. Основная словоформа, соответствующая данной гипотезе, считается "рабочей". Далее выбирается соответствующая ей словоизменительная парадигма, слово включается в ИнфоПортрет.

3.3. Выбор правильной основной формы словосочетаний

В целом метод решения проблемы аналогичен способу снятия омонимии слов, описанному в п.3.2. Мы определяем частоты единственных и множественных чисел отдельных элементов словосочетаний. Элементы словосочетаний будут разными для разных языков, в зависимости от характера появления ошибок при приведении к первой форме. То есть, для русского языка будем запоминать число, единственное или множественное, в котором чаще употребляется прилагательное омонимичного словосочетания, для английского языка – существительное. Определение элемента с наибольшей частотностью позволит выявить, то число, в котором принято употреблять словосочетание в исследуемом массиве данных.

При проведении вычислений гипотезами в формуле (1) принимаем полный набор основных форм словосочетаний с совместным появлением различных сочетаний вариантов форм слов, их составляющих относительно некоторой омонимичной формы словосочетания. По формуле (2) выбираем основную форму словосочетания, содержащую формы слов с наибольшей вероятностью их совместного появления.

4. Полученные результаты

Для оценки результатов приведен ИнфоПортрет, сформированный по запросу "*ПОПЕЧИТЕЛЬСТВО ИЛИ УСЫНОВЛЕНИЕ*". В таблице 4 приведены наиболее значимые его элементы.

Для сравнения также приведены аналоги этих элементов, возникавшие в предыдущей версии (без использования нового подхода). Необходимо отметить, что вследствие изменения способа подсчета частот словосочетаний порядок их ранжирования в ИнфоПортрете также изменяется. Наиболее значимые элементы ИнфоПортрета без использования метода разрешения омонимии, приведены выше, в таблице 3.

ИнфоПортрет, сформированный с применением вероятностного подхода к разрешению омонимии	**ИнфоПортрет, сформированный без применения вероятностного подхода**	**Первые формы словосочетаний, полученных без применения вероятностного подхода**
УСЫНОВЛЕНИЕ	РЕБЕНОК	
РОДИТЕЛЬ	РОДИТЕЛЬ	
ДЕТСКИЙ ДОМ	УСЫНОВЛЕНИЕ	
ПРИЕМНАЯ СЕМЬЯ	ПОПЕЧИТЕЛЬСТВО	
УСЫНОВИТЬ	СОВМЕСТНАЯ СОБСТВЕННОСТЬ	
МОЛОДЫЕ СЕМЬИ	ДИТЯ	**МОЛОДАЯ СЕМЬЯ**
ПРИЕМНЫЕ РОДИТЕЛИ	ПОПЕЧЕНИЕ	**ПРИЕМНЫЙ РОДИТЕЛЬ**
ПОПЕЧИТЕЛЬСТВО	ОСТАТЬСЯ	
РОССИЙСКАЯ ФЕДЕРАЦИЯ	УСЫНОВИТЕЛЬ	
УСЫНОВИТЕЛЬ	ВОСПИТАНИЕ	
ЖИЛОЕ ПОМЕЩЕНИЕ	**ОПЕК**	
ОПЕКА	УСЫНОВИТЬ	
ПОСОБИЕ	ГРАЖДАНСКОЕ СОСТОЯНИЕ	
ПОПЕЧЕНИЕ	АНАЛОГИЧНЫЕ УЧРЕЖДЕНИЯ	
РОДИТЕЛЬСКИЕ ПРАВА	ОПЕКА	**РОДИТЕЛЬСКОЕ ПРАВО**
СИРОТА	УЧРЕДИТЕЛЬ	
ОПЕК	РАСТОРЖЕНИЕ	
ПРИЕМНЫЙ	СЕМЬЯ	
ДЕТЕЙ-СИРОТ	СУДЕБНЫЙ ПОРЯДОК	
ОПЕКУН	ФЕРМЕРСКОЕ ХОЗЯЙСТВО	
ВОСПИТАНИЕ	СУПРУГ	
ПРИЕМНАЯ	БРАК	
ДЕТСКИЙ	ОПЕКУН	
ЕЖЕМЕСЯЧНОЕ ПОСОБИЕ	ИНТЕРЕС	
УЧРЕЖДЕНИЕ	ДЕТСКИЕ ДОМА	
ПАТРОНАТНЫЙ	ПРИЕМНАЯ СЕМЬЯ	
РОССИЙСКИЕ ДЕТИ	АЛИМЕНТЫ	**РОССИЙСКОЕ ДИТЯ**
ФЕДЕРАЦИЯ	СОГЛАСИЕ	
ЖИЛИЩНЫЕ УСЛОВИЯ	ЖЕЛАЮЩИЙ	**ЖИЛИЩНОЕ УСЛОВИЕ**
ПРИЕМНЫЕ ДЕТИ	ПОПЕЧИТЕЛЬ	**ПРИЕМНОЕ ДИТЯ**

Таблица 4. Омонимия в ИнфоПортретах по запросу "Попечительство или усыновление". Жирным шрифтом выделены омонимичные формы слов и словарных пар

Можно отметить, что при использовании нового подхода ранг ошибочного элемента ИнфоПортрета "*ОПЕК*" понизился, в то время как верный элемент "*ОПЕКА"* получил более высокий ранг, чем ошибочный элемент. Кроме того, искаженные по смыслу словосочетания, отображаемые в старом ИнфоПортрете в единственном числе (колонка 3 в таблице 4), в новом ИнфоПортрете приняли правильную форму. Результаты, полученные для "проблемных" словосочетаний ("*CLINTON LEAF*", "*ВООРУЖЕННЫЕ СИЛЫ*" и т.д.), описанных в п.2, были еще более успешны – в ИнфоПортрет они не вошли совсем. Возможно, сыграло свою роль предположение о более редкой встречаемости словосочетаний-омонимов, вынесенное в п.3.2.

Вместе с тем, проявились определенные недостатки метода. Например, омоним "*ОПЕК*" все-таки сохранил достаточно высокий ранг в ИнфоПортрете, пусть и меньший, чем у правильного существительного *"ОПЕКА"*.

5. Заключение

Предложенный метод позволил решить проблемы, связанные с частичной омонимией слов и неверным приведением словосочетаний к первой форме. Данное решение позволило существенно улучшить качество формирования ИнфоПортрета в системе "Галактика-Zoom".

Вместе с тем, следует учитывать, что существующие ограничения реализации данного метода не позволяют создать идеальный алгоритм выявления ошибок. Однако предлагаемые методы позволяют решить поставленные задачи в подавляющем большинстве случаев. Для этого используются как опытные данные, которые предлагается использовать для устранения ошибок, так и концентрация вычислительных ресурсов на работе с наиболее информативными данными, а также статистические характеристики слов и словосочетаний в текстах, уточняющие запросы. Все эти методы позволяют улучшить качество составления ИнфоПортрета.

Список литературы

1. Антонов А.В. Методы классификации и технология Галактика-Zoom // сб. Международный форум по информации. М.: ВИНИТИ, 2003. т.28.
2. Антонов А, Курзинер Е. Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии, Труды Международного семинара Диалог-2002.
3. Антонов А. Информационно-поисковая система Galaktika-ZOOM с элементами анализа на гипермассивах информации // Сб. ВИНИТИ. М.: 2001. №8.
4. Вентцель Е.С. Теория вероятностей. // 7-е изд. стер. М.: Высш. шк., 2001. 575 с.