Московский государственный университет имени М.В. Ломоносова

На правах рукописи

Н.В. Лукашевич

Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа

05.25.05 – Информационные системы и процессы

ДИССЕРТАЦИЯ
на соискание ученой степени
доктора технических наук

Москва 2014

СОДЕРЖАНИЕ

Введение

В настоящее время в связи с огромными объемами электронных документов имеется все возрастающая потребность в обработке неструктурированной текстовой информации, повышению качества и эффективности имеющихся методов обработки текстов. В число активно развивающихся направлений обработки неструктурированной текстовой информации входят такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов, сегментирование документов и многое другое.

Современные информационно-поисковые и информационно-аналитические системы работают с текстовой информацией в широких или неограниченных предметных областях, поэтому характерной чертой современных методов обработки текстовой информации стало минимальное использование знаний о мире и о языке, опора на статистические методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п. В то же время когда подобные операции выполняет человек, ему необходимо выявить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объем знаний о языке, мире, организации связного текста.

Недостаток лингвистических и онтологических знаний (знаний о мире), используемых в приложениях информационного поиска и автоматической обработки текстов, приводит к разнообразным проблемам. Нехватка знаний приводит к нерелевантному поиску в тех случаях, если способы формулировки запросов отличаются от способов описания релевантных ситуаций в документах. Эта проблема усугубляется при обработке длинных запросов, при поиске ответов на вопросы в вопросно-ответных системах.

Так, как показали эксперименты в рамках конференции по информационному поиску TREC и семинаре «Надежный доступ к информации» (Reliable Information Access), проведенном в 2003 году, существуют типы запросов к поисковым системам, которые являются сложными для современных технологий информационного поиска, и, следовательно, качество поиска по этим запросам достаточно низкое. Среди потенциальных методов, которые могли бы улучшить выдачу поисковых систем по таким запросам, указывались методы расширения запросов, в том числе, и с использованием специальных ресурсов, описывающих знания о предметной области.

В последнее время все большее значение приобретают специализированные виды информационного поиска такие, как медицинский, патентный, научный поиск, и роль знаний о предметной области в обеспечении качества работы таких информационных системах, безусловно, значительна. Кроме того, при поиске в отличных от Интернета коллекциях документов, таких, как профессиональные информационные базы, внутрикорпоративные ресурсы, отличающиеся относительно небольшим (по сравнению с Интернет) размером, несоответствие языка запроса и языка документов считается достаточно серьезной проблемой.

Нехватка знаний приводит к снижению качества при автоматической фильтрации и рубрикации документов, к излишним повторам или нарушению связности при автоматическом аннотировании и др.

Одним из типов обычно недостаточно используемых лингвистических знаний в приложениях информационного поиска является неучет структурных свойств связного текста. Как известно, связный текст имеет сложную иерархическую структуру. Одним из существенных проявлений связности текста является так называемая глобальная связность текста, когда в тексте имеется одна главная тема, а вся остальная информация подчинена изложению этой основной темы. Другим способом проявления связности текста является его лексическая связность, когда в тексте содержится

множество близких по смыслу слов и выражений. Между тем подавляющее большинство подходов рассматривает текст как совокупность независимых друг от друга слов (”bag of words”), характеризующихся частотностью встречаемости в документе и коллекции.

В то же время внедрение в современные методы автоматической обработки текстов дополнительных объемов знаний о языке и мире является сложной задачей. Это связано с тем, что такие знания должны описываться в специально создаваемых ресурсах (тезаурусах, онтологиях), которые должны содержать описания десятков тысяч слов и словосочетаний, иметь такие возможности, как логический вывод. При применении таких ресурсов обычно необходимо автоматически разрешать многозначность слов, т.е. выбирать их правильное значение. Кроме того, поскольку ведение любых ресурсов отстает от развития предметной области необходимо развитие комбинированных методов, учитывающих как знания, так и лучшие современные статистические методы обработки текстов.

В настоящее время обсуждаются три основные парадигмы ресурсов, содержащих знания о мире и языке широких предметных областей для использования в информационно-поисковых и информационно-аналитических системах.

Самой первой широко распространенной парадигмой были традиционные информационно-поисковые тезаурусы, разработка и использование которых регламентируется национальными и международными стандартами. Однако такие тезаурусы создавались для ручного индексирования документов людьми-индексаторами, и в последние десятилетия, характеризующиеся резким ростом объемов электронной информации, их роль резко снизилась.

После появления в середине 90-х годов тезауруса WordNet, структура которого представляет собой иерархическую сеть лексикализованных понятий английского языка – синсетов, появились многочисленные работы по использованию такого рода ресурсов в качестве источника

лингвистических знаний в разнообразных приложениях информационного поиска. Однако этот тезаурус создавался для проверки психолингвистической теории, и не учитывает особенностей автоматической обработки текстов, из-за чего имеется много проблем в его использовании в прикладных разработках.

Наконец, современной парадигмой компьютерных ресурсов для приложений информационного поиска являются формальные онтологии, выдвинута концепция Семантической сети (Semantic Web), базирующая на построении онтологических ресурсов большого объема. Однако автоматическую обработку неструктурированных текстов на естественном языке с их неоднозначностью и неточностью трудно проводить с помощью аксиоматизированных теорий, к построению которых стремятся приверженцы формальных онтологий.

Часть исследователей считает, что формальные онтологии должны описывать знания о мире и быть независимыми от конкретного языка. Однако для того, чтобы применить такого рода независимую от языка онтологию в практических задачах информационных технологий, которые во многом связаны с переработкой неструктурированной информации, текстов, необходимо установить отношения между понятиями языково-независимой онтологии и значениями лексических единиц конкретного естественного языка. Кроме того, часть исследователей (см. например, [218]) подвергают сомнению возможность создания большой онтологии совершенно независимо от естественного языка.

Таким образом, при всем обилии научной литературы по вопросам построения информационно-поисковых тезаурусов, тезаурусов типа WordNet, онтологий открытыми остаются следующие вопросы:

- какая модель базы знаний для описания неструктурированной широкой предметной области наиболее оптимальна для того, чтобы, с одной стороны, создать ее в разумные сроки и охватить всю важную для предметной области терминологию, с другой стороны, чтобы созданная

формализованная модель была полезна в широком круге приложений информационного поиска и автоматической обработки текстов,

- как необходимо использовать построенный ресурс для отображения основного содержания текста, с учетом той информации, которая описана в данном ресурсе,

- каковы методы применения полученного ресурса и построенного представления содержания документов в различных задачах информационного поиска, и какого качества решения этих задач можно достигнуть.

Целями исследования, проведенного в диссертации, являются

1) разработка модели представления знаний в предметно-оринтированной базе знаний для описания широких предметных областей с возможностью логического вывода, которая применима для описания многих предметных областей и эффективна при автоматическом построении тематического представления текста, а также в широком круге приложений информационного поиска и автоматической обработки текстов;
2) разработка моделей и алгоритмов для автоматического построения тематического представления текста как иерархической структуры;
3) разработка алгоритмов решения различных задач в сфере информационного поиска и приложений автоматической обработки неструктурированной информации на основе созданных предметно-ориентированных баз знаний и тематического представления текстов;
4) разработка алгоритмов автоматизированного пополнения предметно-ориентированных баз знаний для приложений автоматической обработки неструктурированной информации.

Научная новизна работы. В диссертации разрабатывается система моделей и алгоритмов, направленных на комплексное решение задачи применения знаний о языке и о мире при автоматической обработке текстов для улучшения качества приложений информационного поиска.

Предложена новая формализованная модель базы знаний онтологического типа – лингвистической онтологии, предназначенной для использования в автоматической обработке текстов в широких предметных областях. Модель основывается на сочетании принципов трех различных методологий разработки компьютерных ресурсов:

Модель основывается на сочетании принципов трех различных методологий разработки компьютерных ресурсов:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии создания формальных онтологий.

Предложенная модель позволяет в короткие сроки создавать онтологические ресурсы в неструктурированных предметных областях. При этом созданный ресурс, с одной стороны, будет содержать подробное описание терминологии предметной области, и, с другой стороны, будет иметь внутреннюю структуру, соответствующую современным онтологическим принципам разработки онтологий в виде отличимых понятий и формальных отношений между понятиями, позволяет проведение логического вывода. Особенностью предлагаемого подхода к описанию предметной области является то, что создаваемые предметно-ориентированные базы знаний направлены на эффективное применение в различных задачах информационного поиска, что показано в целом ряде экспериментов.

Предложена модель представления тематической структуры текстов на основе свойств лексической и глобальной связности текста. Предложен и реализован алгоритм автоматического построения тематического представления содержания текстов, которое моделирует основное содержание текста посредством выделения тематических узлов – совокупностей близких по смыслу понятий текста. Выделяются основные

тематические узлы, соответствующие основной теме документа и локальные тематические узлы, соответствующие подтемам документа.

Предложен метод концептуального индексирования документов для информационно-поисковой системы, базирующийся на знаниях, описанных в предметно-ориентированной базе знаний, и построенном тематическом представлении документов. Концептуальный индекс, порождаемый на основе Общественно-политического тезауруса – предметно-ориентированной базе знаний в широкой области общественной жизни современного общества используется в Университетской информационной системе РОССИЯ (www.cir.ru).

Предложен и реализован алгоритм автоматического разрешения лексической многозначности на основе тезаурусных знаний, сочетающий информацию о локальном и глобальном контексте употребления многозначного слова. 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. Метод разрешения многозначности базируется на совокупности различных контекстных признаков и для нахождения их оптимальной комбинации был использован численный метод координатного спуска.

Предложен и реализован алгоритм автоматической рубрикации документов, основанный на использовании тематического представления документов и описании рубрик в виде булевских выражений над понятиями тезауруса и способный обрабатывать тексты различных типов (официальные документы, сообщения информационных агентств, газетные статьи). Система рубрикации легко настраивается на новый рубрикатор и новые типы текстов, рубрицирование можно осуществлять сразу по нескольким рубрикаторам. На основе предложенного метода было реализовано более 20 систем автоматической рубрикации текстов с количеством тематических рубрик от 35 до 3000. Возможности быстрой настройки системы рубрикации на новый

рубрикатор и достигаемый при этом уровень качества рубрикации был продемонстрирован на Российском семинаре по информационному поиску РОМИП в 2007 и 2010 годах. Создание системы рубрикации заняло 8 часов, качество рубрикации было оценено как более чем 70% F-меры.

Предложен и реализован алгоритм автоматического многошагового построения булевского выражения по длинному поисковому запросу на естественном языке, включающий расширение запроса по тезаурусным отношениям, подтвержденным поисковой выдачей. Для обеспечения устойчивости обработки длинного поискового запроса метод построения булевских выражений используется в сочетании с совокупностью различных признаков запроса, документа и коллекции, и для нахождения оптимальной функции соответствия между запросом и документом был использован численный метод координатного спуска.

Предложен и реализован метод автоматического аннотирования отдельного документа, который базируется на тематическом представлении содержания текстов, что позволяет повысить связность создаваемой аннотации. Реализованная система автоматического аннотирования одного документа получила наилучший результат в одной из номинаций на конференции SUMMAC в 1998 году.

Предложен и реализован метод автоматического аннотирования новостного кластера на основе тематического представления кластера и моделировании лексической связности. Показано, что предложенная модель позволяет значительно улучшить связность порождаемой аннотации, а также снизить повторы информации, ухудшающие восприятие порожденного текста человеком.

Предложена и обоснована многофакторная модель извлечения терминов предметной области из текстов. Реализован новый метод автоматизированного извлечения терминов предметной области для пополнения предметно-ориентированной базы знаний. Метод основывается

на вычислении для языковых выражений трех типов статистических характеристик

- характеристик, вычисленных на основе текстовой коллекции предметной области,

- характеристик, вычисленных на основе поисковой выдачи глобальных поисковых систем,

- характеристик, вычисляемых на основе известных терминов предметной области, что очень важно для пополнения предметно-ориентированной базы знаний, учета появляющихся новых терминов в развивающейся предметной области. Для нахождения оптимальной комбинации статистических характеристик для определения терминологичности выражения применяется метод машинного обучения - логистическая регрессия.

Глава 1. Использование знаний в приложениях информационного поиска

1.1. Формальные и лингвистические онтологии

В настоящее время наиболее распространенной формой баз знаний являются базы знаний онтологического типа [64, 67, 186, 239, 249]. Онтологии представляют собой компьютерные ресурсы, содержащие формализованное описание фрагмента знаний о мире. Различные авторы дают разные определения для понятия онтологии [66, 268]. При всем различии к определению онтологии многие авторы соглашаются в наборе основных компонентов онтологии: классы или понятия; атрибуты (свойства); экземпляры (отдельные индивиды), отношения между классами или экземплярами; аксиомы онтологии [36].

Таким образом, формальным определением онтологий может служить следующее:

$$O = <C, E, At, R, A>$$

где C – понятия (классы) онтологии), E – экземпляры онтологии, At – атрибуты понятий и экземпляров онтологии, R – отношения между понятиями, A – аксиомы онтологии.

Термину «онтология» удовлетворяет широкий спектр структур, представляющих знания о той или иной предметной области. В качестве в разной степени формализованных онтологий разными авторами рассматривается множество различных компьютерных ресурсов [156, 216, 268, 312], в том числе и известных задолго до начала исследований по онтологиям таких, как рубрикаторы или тезаурусы.

При этом в некоторых типах онтологий некоторые из вышеперечисленных компонентов онтологий могут быть не определены [239]. Так, рубрикаторы обычно не включают экземпляры и атрибуты, т.е. распространенной формальной моделью рубрикаторов является модель вида:

$$O = <C, \varnothing, \varnothing, R, A> = < C, R, A >$$

Наиболее формализованные онтологии представляют собой **логические теории**, построенные на произвольных логических утверждениях о понятиях – аксиомах. Для описания таких формальных онтологий применяются различные логики (дескриптивные логики, модальные логики, логика предикатов первого порядка) и различные языки описания онтологий DAML+OIL, OWL, CycL, Ontolingua.

Онтологии, такие, как тезаурусы, рубрикаторы, понятия которых не определяются полностью в терминах формальных свойств и аксиом, иногда называются **легкими онтологиями** (lightweight ontologies) [60]. Дж. Сова (http://www.jfsowa.com/ontology/ontoshar.htm) называет такие онтологии **терминологическими онтологиями**.

Разработчики онтологий по-разному трактуют взаимоотношения между онтологией и естественным языком. Некоторые исследователи трактуют онтологию как структуру, независимую от естественного языка, другие – как структуру, независимую от *конкретного* естественного языка, третьи вводят элементы языкового лексикона в формальное определение онтологии [75, 82, 121,122, 151].

Обсуждая вопросы построения онтологий, многие исследователи подчеркивают значимость текстов как источника знаний о предметной области. Так, в работе [200] указывается, что тогда как небольшие онтологии могут быть построены методом сверху-вниз, разработка подробных онтологий для реальных приложений – нетривиальная задача. Более того, во многих предметных областях, знание, нужное для распространения и интеграции, содержится в основном в текстах. Из-за внутренних свойств человеческого языка, непростой задачей является связать знания, содержащиеся в текстах, с онтологиями, даже если бы они были построены для данной предметной области. Авторы вышеуказанной работы делают вывод, что такие однозначные и последовательные концептуальные модели

играют менее значительную роль в распространении знаний, чем предполагают сторонники формального онтологического подхода.

Еще одной важной проблемой построения онтологий, частично связанная с естественным языком, является проблема понятности онтологии для пользователей так, чтобы она могла правильно применяться и интерпретироваться [19, 51, 217]. На основе спецификаций и документации онтологии пользователи должны правильно интерпретировать семантику всех ее элементов. Кроме того, как показывает практика, далеко не всякий специалист в предметной области может хорошо разбираться в формальных онтологических спецификациях. Чем больше степень формализованности онтологии, тем труднее ее понять пользователю.

Вместе с тем имеется немало подходов к построению онтологий, в которых компоненты лексикона предметной области непосредственно вводятся в формальное определение онтологии [121, 122, 309]. Так, одной из известных формальных моделей онтологии является модель, описанная в [121]:

$$O = <L, C, F, G, H, R, A>$$

где $L = L_C \cup L_R$ – словарь онтологии, содержащий набор лексических единиц (знаков) для понятий L_C и набор знаков для отношений L_R;

- C – набор понятий онтологии;
- F и G связывают наборы лексических единиц $\{l_j\} \subset L$ с наборами понятий и отношений данной онтологии;
- H – фиксирует таксономический характер отношений (связей), при котором понятия онтологии связаны нерефлексивными, ациклическими, транзитивными отношениями $H \subset C \times C$;
- R – обозначает нетаксономические отношения между понятиями онтологии,
- A – набор аксиом онтологии.

Вместе с тем, даже в таких подходах, рассматривающих лексикон естественного языка как один из компонентов онтологической модели, ничего не говорится о методах установления соответствия между совокупностью лексических значений текстов предметной области и онтологии, лексические выражения представлены как вспомогательные элементы, называющие понятия и отношения онтологии.

Однако в установлении взаимоотношений между понятиями и словами и выражениями естественного языка имеется много проблем, начиная с того, как ввод нового понятия в онтологию связан с существующими языковыми выражениями. Кроме того, стремление к четкой формализации отношений между понятиями в онтологии чрезвычайно трудно соблюсти в ситуации, когда необходимо создавать сверхбольшие ресурсы, и, кроме того, приводит к проблемам при установлении связей «понятие – языковое выражение» [77, 125, 152].

Поэтому значительно большее распространение в приложениях автоматической обработки текстов получили вышеупомянутые "легкие" онтологии. Так, большое количество широкоизвестных медицинских онтологических ресурсов представляет собой тезаурусы, не обладающие высокой степенью формализации своей структуры [55].

Тезаурусы представляют собой так называемые **лингвистические онтологии**, т.е. онтологии, опирающиеся в своем построении на значения реально существующих выражений естественного языка. Наиболее известными типами тезаурусов, обсуждаемыми в качестве источников знаний для приложений обработки неструктурированной информации, являются информационно-поисковые тезаурусы и тезаурусы типа WordNet, структура которых будет рассмотрена ниже.

1.1.1. Информационно-поисковые тезаурусы

Информационно-поисковый тезаурус (в соответствии с определениями стандартов) – это нормативный словарь терминов на естественном языке,

явно указывающий отношения между терминами и предназначенный для описания содержания документов и поисковых запросов [83, 84, 223, 245].

Основными целями разработки традиционных информационно-поисковых тезаурусов являются следующие:

- обеспечение перевода естественного языка документов и пользователей на контролируемый словарь, применяемый для индексирования и поиска;
- обеспечение последовательного использования единиц индексирования [243-246];
- описание отношений между терминами;
- использование как поискового средства при поиске документов.

Основной единицей тезаурусов являются термины, которые разделяются на дескрипторы (=авторизованные термины) и недескрипторы (=аскрипторы). По своей сути дескрипторы однозначно соответствуют понятиям предметной области [223].

Отношения между дескрипторами обычно разделяются на два типа: иерархические и ассоциативные. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные.

По ГОСТУ 7.25-2001 [245] иерархические отношения обладают свойствами транзитивности и антисимметричности, которые могут быть использованы при избыточном индексировании в интересах повышения эффективности информационного поиска. Предпочтительно указывать связи между дескрипторами как отношения иерархического вида, если они обладают этими свойствами. Применяемые в ИПТ иерархические отношения могут дифференцироваться на отдельные виды.

Основным иерархическим отношением, используемым в информационно-поисковых тезаурусах, является родовидовое отношение *выше-ниже*. Родовидовая связь устанавливается между двумя дескрипторами, если объем понятия нижестоящего дескриптора входит в

объем понятия вышестоящего дескриптора. Также в качестве иерархического отношения в информационно-поисковых тезаурусах может устанавливаться отношение *часть-целое*.

Отношение ассоциации является неиерархическим и ассоциативным. Основное назначение установления ассоциативных отношений между дескрипторами информационно-поискового тезауруса – указание на дополнительные дескрипторы, полезные при индексировании или поиске.

Основной целью разработки традиционных информационно-поисковых тезаурусов является использование их единиц (дескрипторов) для описания основных тем документов в процессе ручного индексирования. Поэтому важно, чтобы набор дескрипторов информационно-поискового тезауруса позволял описывать тематику документов предметной области [106].

При этом сам процесс индексирования по такому тезаурусу базируется на лингвистических, грамматических знаниях, а также знаниях о предметной области, которые имеются у профессиональных индексаторов текстов. Индексатор сначала должен прочитать текст, понять его и затем изложить содержание текста, пользуясь дескрипторами, указанными в информационно-поисковом тезаурусе. Индексатор должен хорошо понимать всю терминологию, использованную в тексте, – для описания основной темы текста ему понадобится значительно меньшее количество терминов.

Таким образом, формальную модель информационно-поискового тезауруса можно представить следующим образом:

$$ИПТ = <D_{th},\ T,\ R_H,\ R_A,\ A_T>$$

где D_{th} – набор дескрипторов предметной области, соответствующий понятиям данной предметной области, индекс $_{th}$ означает в данном случае тот, факт что разработчики информационно-поисковых тезаурусов включают в состав дескрипторов термины предметной области, которые необходимы для выражения основных тем документов этой ПО [106];

T – набор терминов предметной области: $D \subset T$; R_H – иерархические отношения информационно-поискового тезауруса; R_A – ассоциативные отношения информационно-поискового тезауруса; A_T – аксиомы транзитивности иерархических отношений.

Отметим, что описанная в национальных и международных стандартах модель информационно-поискового тезауруса предназначена для его использования в процессе ручного, экспертного анализа документов [83, 223]. Информационно-поисковый тезаурус, предназначенный для автоматической обработки текстов, должен содержать значительно больше информации о структуре и языке предметной области. Кроме того, отношения между терминами, указанные в тезаурусе, должны быть значительно более формализованы для использования их в автоматических режимах. Если применять традиционные информационно-поисковые тезаурусы в автоматической обработке текстов, то возникает ряд существенных проблем.

1.1.2. Тезаурусы типа WordNet

Лингвистический ресурс WordNet разработан в Принстонском университете США. WordNet относится к классу лексических онтологий, свободно доступен в Интернет, и на его основе были выполнены тысячи экспериментов в области информационного поиска [138]. WordNet версии 3.0 охватывает приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий, или совокупностей синонимов (synset); общее число пар лексема-значение насчитывает 200 тысяч. В разных странах предприняты усилия по созданию ресурсов для своих языков по модели WordNet [12, 13, 211, 212, 231, 308].

Основным отношением в WordNet является отношение синонимии. Наборы синонимов – синсеты – основные структурные элементы WordNet. Понятие синонимии базируется на критерии, что два выражения являются

синонимичными, если замена одного из них на другое в предложении не меняет значения истинности этого высказывания.

Именно определение синонимии в терминах заменимости делает необходимым разделение WordNet на отдельные подструктуры по частям речи. В состав словаря входят лексемы, относящиеся к четырем частям речи: прилагательное, существительное, глагол и наречие. Лексемы различных частей речи хранятся отдельно и описания, соответствующие каждой части речи, имеют различную структуру.

Синсет может рассматриваться как представление лексикализованного понятия (концепта) английского языка. Авторы считают, что синсет существительных представляет понятия существительных, глаголы выражают глагольные концепты, прилагательные – концепты прилагательных и т.п. Кроме того, авторы считают, что такое разделение соответствует психолингвистическим экспериментам, что представление информации о прилагательных, существительных, глаголах и наречиях устроено в человеческой памяти по-разному.

Большинство синсетов снабжены толкованием, подобным толкованиям в традиционных словарях, — это толкование рассматривается как одно для всех синонимов синсета. Если слово имеет несколько значений, то оно входит в несколько различных синсетов.

Каждая часть речи в WordNet имеет свой набор отношений. В различных компьютерных приложениях чаще всего используются существительные, между которыми установленя отношения синонимии, антонимии, гипонимии (гиперонимии), меронимии (*часть-целое*).

Основным отношением между синсетами существительных является родовидовое отношение, при этом видовой синсет называется гипонимом, а родовой — гиперонимом. Это транзитивное иерархическое отношение, которое может быть также названо isA-отношение. Синсет X называется гипонимом синсета Y, если носители английского языка считают нормальными предложения типа «An X is a (kind of) Y”.

Таким образом, отношения между синсетами образуют иерархическую структуру. При построении иерархических систем на базе родовидовых отношений обычно предполагается, что свойства вышестоящих понятий наследуются на нижестоящие – так называемое свойство наследования. Таким образом, существительные в WordNet организованы в виде иерархической системы с наследованием; были сделаны систематические усилия, чтобы для каждого синсета найти его родовое понятие, его гипероним.

Появление WordNet и возможность его свободного использования вызвали большое число исследований по применению этого тезауруса в самых различных приложениях автоматической обработки текстов. Большое количество экспериментов привело к массовому выявлению и обсуждению проблем и недостатков WordNet, препятствующих его эффективному применению.

Так, при разработке WordNet был выдвинут принцип раздельного описания разных частей речи. Между различными частями речи, имеющими одинаковое значение, не было установлено никаких отношений. Это вызывало серьезные проблемы в приложениях, поскольку одно и та же мысль могла быть выражена разными частями речи [35]. Кроме того, в различных языках для выражения одной и той же идеи могут использоваться лексемы разных частей речи. Поэтому иерархии синсетов, построенные на основе конкретных частей речи, становятся в большой мере зависимыми от естественного языка разработки, поскольку в некотором естественном языке может не оказаться возможности выразить некоторое понятие той или иной частью речи. Начиная с версии WordNet 2.0, в ресурс были введены отношения между однокоренными синсетами, относящимися к разным частям речи и связанными между собой по смыслу [139].

Другой проблемой, вызвавшей серьезное обсуждение среди исследователей, стало описание значений многозначных слов в WordNet. Во многих работах признается, что различия значений в WordNet слишком

тонки для таких компьютерных приложений, как машинный перевод, информационный поиск, классификация текстов, вопросно-ответные системы и др. В [33] было показано, что среднее количество значений в WordNet больше, чем в традиционных лексикографических словарях. Эти проблемы привели к постановке вопроса о том, каким образом и какие типы значений многозначного слова могут быть объединены («кластеризованы») [32, 163] для целей работы в приложениях автоматической обработки текстов, когда для значений многозначного слова из кластера можно не делать различий, и это не приведет к снижению качества работы этого приложения.

Современные версии WordNet содержат для каждого многозначного слова указание на самое частотное значение по корпусу SemCor [94], что дает возможность в случае проблем при процедуре автоматического разрешения многозначности выбирать это самое частотное значение.

Одной из серьезных проблем WordNet, препятствующей его использованию в приложениях, является нехватка разнообразных отношений между синсетами. В частности, исследователями широко обсуждалась так называемая «теннисная проблема»: принадлежащие одной предметной области, сфере деятельности, ситуации синсеты оказываются очень далеко друг от друга в структуре WordNet [138]. Отсутствие такого рода отношений оказывает серьезное негативное воздействие на использование WordNet в автоматических процедурах разрешения лексической многозначности, вызывает проблемы в информационном поиске.

Формальную модель ресурса типа WordNet можно представить следующим образом:

$WN = <LC_{n,adj,v,adv}, R_{n,adj,v,adv,} S, T, M, A_n >$

где $LC_{n,adj,v,adv} = \{LC_n, LC_{adj}, LC_v, LC_{adv}\}$ – совокупность лексикализованных понятий-синсетов, сгруппированных по разным частям речи (существительные, прилагательные, глаголы и наречия); синсет

представляется собой одну лексему (слово в определенном значении) или совокупность синонимичных лексем,

- $R_{n,adj,v,adv}$= {R_n, R_{adj}, R_v, R_{adv}} – наборы отношений синсетов, различающиеся для разных частей речи,

- T – текстовые выражения (слова и словосочетания), описанные в ресурсе,

- S – отношения между текстовыми выражениями и синсетами,

- M – совокупность неоднозначных текстовых выражений: $M \subset T$,

- A_n – аксиомы транзитивности и наследования, индекс n отражает тот факт, что аксиомы обсуждаются и используются в подавляющем большинстве случаев только для синсетов существительных.

В результате рассмотрения структурных особенностей информационно-поисковых тезаурусов и тезаурусов типа WordNet, можно сделать следующие выводы о сходстве и различии используемых моделей представления знаний в этих тезаурусах.

Наиболее бросающееся в глаза различие состоит в том, что информационно-поисковые тезаурусы описывают определенную предметную область, а WordNet содержит информацию о значениях общей лексики языка. Однако это различие не является принципиальным, поскольку можно строить тезаурусы типа WordNet и для конкретных предметных областей [15, 24, 124, 131, 175, 176]. Более значимые различия имеются в выборе единиц тезаурусов.

В информационно-поисковых тезаурусах имеется множество ограничений на включение в тезаурус языковых единиц: дескрипторы должны быть четко отделены по смыслу друг от друга, многозначность языковых единиц практически не представлена, ограничивается глубина иерархий и т.д. Это приводит к возникновению существенного расхождения между единицами тезауруса и языковыми единицами, упоминаемыми в текстах предметной области. В тезаурусах типа Wordnet такой разницы нет:

если существует слово или выражение с определенными значениями, то оно включается в тезаурус в соответствующем количестве значений.

Существенно различным является подход к включению в эти два типа тезаурусов словосочетаний. В информационно-поисковых тезаурусах имеется достаточно подробный перечень правил, которыми должен руководствоваться разработчик тезауруса при вводе в тезаурус многословных дескрипторов. Разработчики WordNet заявляют о необходимости того, чтобы словосочетание было «лексикализовано» без уточнения критериев, а это, в свою очередь, приводит к тому, что ввод новых словосочетаний в WordNet, а особенно в тезаурусы типа Wordnet, создаваемые для других языков, серьезно ограничивается.

Если сравнивать систему отношений в стандартных информационно-поисковых тезаурусах и тезаурусах типа WordNet, то, прежде всего, нужно брать для сравнения отношения между синсетами существительных WordNet, поскольку дескрипторы информационно-поисковых тезаурусов – это обычно существительные и группы существительного.

Оба типа тезаурусов имеют небольшой набор отношений, что, несомненно, объясняется разнообразием описываемых сущностей. При этом, однако, в наборе отношений информационно-поискового тезауруса имеется отношение ассоциации, которое при всей высказанной по поводу его критики [102, 201, 202] позволяет лучше описать отношения между сущностями предметной области, чем отношение *часть-целое* в версии WordNet и *антонимии*.

В последнее время в ряде работ отмечается, что и разработчики информационно-поисковых тезаурусов и разработчики ворднетов включают в свои тезаурусы более разнообразные наборы отношений между единицами [34, 188].

1.2. Методы применения лингвистических онтологий в приложениях обработки неструктурированной информации

1.2.1. Автоматическое концептуальное индексирование на основе информационно-поисковых тезаурусов

Онтологии и тезаурусы могут применяться в так называемом концептуальном индексировании текстов в рамках информационных систем. В концептуальном индексе, в отличие от пословного индекса, слова-синонимы должны быть соединены вместе, в один элемент индекса, а разные значения многозначных слов разделены [31, 114, 135, 221, 250].

Поскольку основными элементами информационно-поискового тезауруса являются термины, описанные как дескрипторы и аскрипторы, может показаться, что достаточно просто осуществить автоматическое индексирование по информационно-поисковым тезаурусам путем простого сопоставления дескрипторов и аскрипторов с документами.

Однако для большинства документов такое автоматическое сопоставление не сможет отразить основное содержание документа:

- важные термины документа могут быть не найдены в тезаурусе, поскольку выражены в нем несколько иначе;
- менее значимые термины найдут прямое отражение в тезаурусе и выйдут на первый план и т.п.;

В работе [167] приводятся данные, полученные на основе 587 документов, проиндексированных вручную дескрипторами тезауруса EUROVOC. Только 31% документов явно содержит в тексте дескрипторы, приписанные документу индексаторами. При этом в 9 из 10 случаев дескрипторы, найденные в тексте документа, не приписаны индексаторами.

Поэтому исследуются более сложные методы автоматизации индексирования по информационно-поисковым тезаурусам.

Одним из подходов для автоматизации индексирования по традиционным информационно-поисковым тезаурусам является подход, основанный на правилах. Такой подход к автоматическому индексированию был реализован по тезаурусу EUROVOC [265], для чего было создано около 40 тысяч правил [79].

В качестве других подходов автоматизации индексирования используются статистические методы и машинное обучение [144, 167, 195]. При таких подходах процесс автоматического приписывания дескрипторов информационно-поискового тезауруса полнотекстовым документам включает две стадии.

На первой стадии (этап обучения) на основе документов, вручную проиндексированных индексаторами, устанавливается соответствие между словами, встретившимися в тексте документа, и приписанными дескрипторами тезауруса. На второй стадии (собственно, индексирование) для каждого слова документа проверяется, каким дескрипторам тезауруса оно соответствует. Если такие дескрипторы имеются, то слово добавляет к весу дескриптора для данного текста натуральный логарифм веса, полученного на первом этапе. После обработки всех слов текущего текста получается суммированный вес дескрипторов тезауруса.

Понятно, что применение таких методов для автоматического индексирования по традиционым информационно-поисковым тезаурусам требует создания большой обучающей выборки, представляет собой по сути классификацию текстов на большое количество классов (по числу дескрипторов тезауруса), с чем в настоящее время системы машинного обучения справляются не очень хорошо. К тому же серьезным фактором, затрудняющим обучение, является субъективность ручного индексирования.

Установление соответствий между отдельными словами и дескрипторами тезауруса дает возможность использовать эти корреляции для обработки поисковых запросов пользователя.

Так, в работах [164, 165] описываются эксперименты по автоматическому расширению свободного запроса пользователя дескрипторами двуязычного тезауруса по социальным наукам [182], которые проводились на двуязычной коллекции немецких и английских документов по общественным наукам. База включает в себя более 150 тысяч немецких документов и 26 тысяч — английских. Документы реферативного характера содержат заголовок публикации, реферат и дескрипторы Тезауруса по общественным наукам, приписанные индексаторами. Эксперименты выполнялись в рамках предметно-ориентированного задания форума по многоязыковым информационным системам CLEF [91].

Для каждого слова запроса выявлялись два наиболее коррелирующих с этим словом дескриптора тезауруса и добавлялись в запрос. Было получено, что в этом случае показатель средней точности поиска для 25 запросов возросла с 45.5% до 51.4%, т.е. более чем на 13% для немецкого языка, и с 45.1% до 48.2% для английского языка.

1.2.2. Автоматическое разрешение лексической многозначности

В случае если в используемом для концептуального индексирования ресурсе (например, тезаурусе типа WordNet) представлены разные значения многозначных слов, то важным является обеспечение качественной процедуры разрешения лексической многозначности, т.е. автоматического выбора между разными значениями слов и словосочетаний, перечисленных в лингвистическом ресурсе.

Применение тезаурусов и онтологий в информационном поиске требует высокого качества разрешения многозначности слов. Так, в работе [178] обосновывалось, что для того, чтобы в информационном поиске мог проявиться положительный эффект от разрешения лексической многозначности, точность разрешения многозначности должна быть не меньше 90%, в работе [63] на основании результатов проведенных

экспериментов указывается необходимая величина точности разрешения многозначности – 70%.

В последние годы проблема разрешения лексической многозначности стала исследоваться как отдельная задача. С 1998 года 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. 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 автоматического 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 [90, 187].

Максимальное качество, которое может достигнуть система автоматического разрешения многозначности, ограничивается согласием между ручной разметкой, сделанной разными экспертами. В настоящее время, согласие между экспертами достигает 95% и выше для четко различимых значений. Для многозначных слов со значениями, близкими по смыслу, согласие между экспертами может составлять 65-70%.

Нижняя граница качества разрешения многозначности определяется на основе случайно выбранного значения (предполагается равновероятность значений) или наиболее частотного значения (предполагается, что вероятность одного значения многократно превышает вероятности других значений). Также в качестве базового метода для сравнения используется так

называемый метод Леска, который основан на сопоставлении словарных толкований слов, упомянутых в анализируемом фрагменте текста [90, 98].

Для автоматической обработки текстов наиболее существенны результаты, которые достигаются современными системами в задаче разрешения всех многозначных слов текста. Для тестирования задачи «все слова текста» на конференции Senseval-3 использовались три текста: две статьи из Wall Street Journal и фрагмент из Брауновского корпуса – общий объем 5000 слов [90, 187]. Всего для тестирования использовались 2081 слов. Семантическая разметка текстов проводилось по набору значений тезауруса WordNet. Если в WordNet не было подходящего значения, то проставлялась помета U.

По результатам конференции SENSEVAL-3 для английского языка в задаче разрешения многозначности для всех слов текста точность лучшей системы составляет 65.2% [187].

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 [187].

Согласие между лексикографами-аннотаторами значений достигало – 72,5%. Наибольший процент разногласий по разметке значений был связан с небольшим набором трудных слов, например, *national*.

Рассмотрим алгоритмы разрешения лексической многозначности на основе структуры тезауруса английского языка WordNet.

Одним из классов предлагаемых методов является оценка семантической близости контекста вхождения того или иного многозначного

термина к 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 [96] предполагается, что два значения тем семантически ближе, чем короче связывающий их путь. Упор делается на отношения гипонимии-гиперонимии и взвешивается длина пути относительно всей глубины таксономии (*D*):

$$Sim_{LC}(C1, C2) = -\log(PathLen(C1, C2)/2D) \qquad (1.1)$$

В работе [76] предполагается, что два синсета семантически близки, если соединены достаточно коротким путем, который имеет малое количество перегибов:

$$Sim_{HS}(C1, C2) = c_0 - PathLen - k \cdot d \qquad (1.2)$$

где d – количество перегибов на протяжении пути; c_0 и k – константы. Если такого пути не существует, то $Sim_{HS}(C1, C2) = 0$. В экспериментах использовались значения констант $c_0 = 8$, $k = 1$, максимальная длина пути 5 шагов.

В ряде работ концептуальное расстояние между синсетами учитывает большее число параметров. Так, для подсчета концептуального расстояния в работе [5, 6] вводится понятие *концептуальной плотности* и формула ее вычисления, которая, по мнению авторов, наилучшим способом описывает близость между словами. В формуле учитываются следующие параметры:

- длина самого короткого пути в иерархии;
- глубина в иерархии;
- плотность понятий в иерархии;
- число концептов.

Формула вычисления концептуальной плотности выглядит следующим образом:

- с-корень (вершина);

- nhyp – число гипонимов в вершине;
- h – высоту иерархии;
- m – число слов из контекста, которые попали в иерархию.

Тогда формула, которая вычисляет плотность (1.3).

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^i}{\sum_{i=0}^{h-1} nhyp^i} \tag{1.3}$$

Эти формулы автор пытался улучшить опытным путем, вводя параметры, и смотря, при каких значениях формула дает наилучшие результаты. В итоге выбор был остановлен на формуле (1.4).

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} (nhyp + \beta)^{i^{\alpha}}}{descendants_c} \tag{1.4}$$

где $descendants_c = \sum_{i=0}^{h-1} nhyp^i$ (1.5)

Другим направлением выбора значения многозначного слова на основе близости контекста в тексте и окружения слов в тезаурусе являются подходы, основанные на оценке так называемого информационного содержания.

Ф. Резник [171] 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 C_1 вид для C_2, то $P(C_1) =< P(C_2)$. 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:

$$IC\,(C) \;=\; -\,log\,(P(C))\,. \tag{1.6}$$

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.

$$Sim_{Rz}(C_1, C_2) = IC(LCS(C_1, C_2)) \quad (1.7)$$

Авторы работы [87] развивают формулу (1.7) следующим образом:

$$Sim_{JC}(C_1, C_2) = 2IC(LCS(C_1, C_2)) - (IC(C_1) + IC(C_2)), \quad (1.8)$$

т.е. 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 [161] описывается тестирование ряда предложенных на базе WordNet метрик на материалах конференции Senseval-2. Для 1723 многозначных существительных коллекции метрики применялись в контексте длиной одно слово. Например, для выражения *Plant with flowers,* по этим мерам вычислялось сходство существительных *plant и flower.* Лучший результат был получен для метрики, предложенной в работе [87], и составил 39% точности.

В работе [213] предлагается алгоритм разрешения лексической многозначности на основе разметки предметных областей Wordnet [123], при которой большинство синсетов тезауруса 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 Factotum, т.е. к словам, не относящимся к конкретным предметным областям таким, как *быть, начинаться, человек*.

Подход к разрешению многозначности на основе содержания целого текста тестируется в работе [52]. 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, что в тексте встречается только одно значение многозначного слова.

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, независимо 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 корпуса Semcor оценивается как 62.09%.

Авторы работы [137] используют известный алгоритм PageRank [158] 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. В результате 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%.

Таким образом, достигнутые показатели разрешения многозначности для задачи «все слова текста», которые собственно и является базой для последующей обработки текста, не кажутся достаточно высокими, поскольку не достигают и 70% точности.

1.2.3. Тезаурусы типа WordNet в информационном поиске

В работах [207, 208] описываются эксперименты по интеграции WordNet в поиск по векторной модели.

Векторная модель информационного поиска предполагает независимость употребления слов в тексте и представляет поисковый запрос и документ в виде векторов слов с весами [177]:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$$

где d_j – векторное представление j-го документа, w_{ij} – вес i-го слова в j-м документе, n – общее количество различных слов во всех документах коллекции.

Основным способом вычисления весов слова w_{ij} в векторе документа является мера tf.idf (term frequency – inverse document frequency, частота терма - обратная частота документа), т.е. вес определяется как произведение функции от количества вхождений терма в документ и функции от величины, обратной количеству документов коллекции, в которых встречается этот терм. Idf часто вычисляется по следующей формуле [129]:

$$Idf_{wj} = \log\left(\frac{N}{n_j}\right),$$

где N – число документов в коллекции, n_j – число документов, в которых встретилось слово w_j.

Для определения сходства между векторами запроса и документа используется так называемая косинусная мера.

$$Sim\ (q, d) = \frac{\sum w_{qi} \cdot w_{di}}{\sqrt{\sum w_{qi}^{\ 2}} \cdot \sqrt{\sum w_{di}^{\ 2}}}$$

Таким образом, теперь соответствие запроса документу измеряется конкретным числом, и все документы могут быть упорядочены в выдаче поисковой системы по этому числу.

Целью экспериментов в работах [207, 208] была попытка выполнить поиск документов на основе не отдельных слов, а значений WordNet. Для каждого документа сначала выполняется процедура разрешения многозначности существительных, которая выдает единственный синсет, и в результате которой каждому тексту ставится в соответствие вектор синсетов WordNet. После того, как вектор создан, с ним могут выполняться такие же операции, как и с пословными векторами.

Эффективность использования векторов синсетов сравнивалась с эффективностью информационного поиска на основе стандартного вектора слов. В стандартном прогоне и документы, и запросы представляются как вектора лемм всех значимых слов. Для экспериментов было использовано 5 разных коллекций документов (компьютерная область, медицинская область, газетные статьи и др.), и для каждой коллекции было выполнено более 30 различных запросов.

Эффективность информационного поиска оценивалась на основе меры средней точности (average precision), которая усредняет точность при выдаче каждого из K релевантных документов.

Данная мера считается следующим образом. Точность на уровне i-го релевантного документа *prec_rel*(*i*) равна *precision* (*pos*(*i*)), если релевантный

документ находится в результатах запроса на позиции *pos*(*i*). Если i-й релевантный документ не найден, то *prec_rel*(*i*)=*0.* Средняя точность для заданного запроса равна среднему значению величины *prec_rel*(*i*) по всем k релевантным документам:

$$\text{AvgPrec}= \frac{1}{k} \sum_{i=1,k} prec\_rel(i)$$

Было показано, что возникает значительное ухудшение качества поиска для векторов, включающих синсеты (от 6.2 до 42.3%), что связано с тем, что часто возникают несоответствия между значением слова, выбранным в запросе, и значением того же слова, выбранным в релевантном документе.

В другой группе экспериментов по использованию WordNet в информационном поиске исследовалась возможность расширения запроса синонимами или другими словами, связанными со словами запроса отношениями, описанными в WordNet. В таких экспериментах нет необходимости выбора единственного значения слова, что в случае ошибки приводит к серьезному ухудшению результатов поиска. Сначала для каждого слова запроса, частотность которых меньше некоторого числа N, и каждого синсета для значений этого слова извлекается список близких по WordNet слов. Те слова, которые встретились, по крайней мере, в двух таких списках, добавляются к исходному запросу. Максимальное улучшение, которое удалось получить – 0.7% средней точности, что не является статистически значимой величиной (N=5%, расстояние – 2, вес расширения w=0.3).

Основные выводы автора работы [208] заключались в том, что для успешного применения WordNet в информационном поиске необходимо значительно улучшить эффективность автоматического расширения лексической многозначности.

В рамках европейского проекта Meaning, который является развитием проекта EuroWordNet, голландская компания Irion Technologies разработала технологию концептуального индексирования TwentyOne, комбинирующую лингвистический и статистический подходы [213, 214]. Авторы разработки считают, что неудачи с использованием WordNet в информационно-поисковых приложениях связаны с трудностями встраивания такого рода лингвистических ресурсов в приложения, оптимального использования содержащейся в ворднетах информации.

Основой технологии является статистическая машина поиска, базирующаяся на стандартной векторной модели и обеспечивающая быстрый поиск документов. Лингвистические технологии используются в двух ролях:

- максимизация полноты выдачи статистической машины за счет синонимии ворднетов;
- максимизация точности выдачи за счет сравнения запросов с конкретными фразами документов, а не с целыми документами.

Фраза представляет собой именную группу (noun phrase). Каждая фраза ассоциируется с отдельными словами, определенной комбинацией слов, а также комбинацией частей слов. Система TwentyOne использует совокупность факторов для сравнения запроса с фразами текста, например: число совпадающих синсетов между запросом и каждой фразой, степень нечеткого сопоставления между запросом и каждой фразой, степень деривационного несовпадения, слитного-раздельного написания и др.

При обработке запроса сначала с помощью векторной модели находятся документы, соответствующие запросу. Затем выданные документы переранжируются так, что сначала выдаются документы, которые имеют наибольшее совпадение по синсетам фраз с запросом. Среди документов, имеющих одинаковое количество сопоставленных синсетов между собственными фразами и запросом, первыми выдаются наиболее похожие по

конкретному набору слов. Вес документа по векторной модели используется, если вес по фразам текста получился одинаковым.

Разрешение многозначности в данной системе делается на основе технологии, описанной в [123], и базируется на разметке предметных областей wordnet. В результате проведенного тестирования авторы работ [213, 214] делают вывод о полезности тезаурусов типа WordNet для информационного поиска, однако из-за специфической процедуры формирования тестового набора запросов трудно оценить, насколько этот вывод обоснован в данных экспериментах.

Вопрос о том, улучшит ли разрешение многозначности слова поиск по словам в правильном значении, остается дискуссионным. Некоторые авторы (Voorhees, Sanderson) полагают, что если запрос однозначно определяет значение многозначного слова в своем составе, то и в найденных документах, это слово окажется в окружении тех же слов запроса, и тем самым с большой вероятностью будет употребляться в том же значении.

Если же выполняется автоматическая процедура разрешения лексической многозначности, то ошибки в работе этой процедуры могут привести к значительному снижению качества информационного поиска, как это и было показано в экспериментах H.Voorhees [207-208]. В работе [178] автор вводит в коллекцию искусственную многозначность и тем самым может контролировать процент ее ошибочного разрешения. В исследовании было показано, что при качестве разрешения многозначности хуже 90% эффективность информационного поиска начинает резко снижаться.

В исследовании [63] авторы ставят перед собой два вопроса:

1) Абстрагируясь от проблемы разрешения многозначности, какой потенциал несет использование ресурсов типа WordNet для информационного поиска. Такой эксперимент можно выполнить, если сделать вручную разрешение лексической многозначности запросов и документов;

2) Если эффективность использования WordNet для коллекции с разрешенной многозначностью известна, то можно измерить чувствительность качества информационного поиска к ошибкам разрешения многозначности, искусственно внося некоторый процент ошибок в разметку по значениям.

Исследования выполнялись на корпусе SemCor, размеченного значениями WordNet. Были выбраны 171 текстовых фрагментов со средней длиной 1331 слова на документ. Для каждого текста была написана краткая аннотация длиной от 4 до 50 слов, в среднем 22 слова на документ. Эти аннотации использовались как запросы по текстовой коллекции, т.е. был ровно один релевантный документ на запрос. Аннотации также были размечены по значениям WordNet. На основе стандартного списка стоп-слов английского языка был также автоматически порожден список стоп-синсетов.

В экспериментах использовалась векторная модель в версии информационно-поисковой системы SMART [177] и три типа векторов: исходные слова документа, значения слов, соответствующие словам документа, и синсеты WordNet, соответствующие словам документа (в последнем случае фактически производится дополнение документа синонимами слов). В процессе эксперимента выяснялось, какой процент документов был возвращен на первом месте в выдаче. Эксперименты показали, что стандартная векторная модель дает 48% первых релевантных документов, индексирование по значениям слов – 53.2% и индексирование по синсетам – 62%.

Внесение ошибок разрешения многозначности в индексирование по синсетам показало, что 10% ошибок не влияет на качество поиска, что находится в соответствии с работой [178]. При этом выяснилось, что при уровне 30% ошибок качество поиска превосходит поиск по стандартной модели SMART (54.4%). Таким образом, авторы делают вывод, что если

выполнять разрешение многозначности с точностью больше 70%, то это даст преимущество по сравнению с пословными векторными моделями.

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 которой является применение алгоритмов разрешения многозначности в рамках задачи информационного поиска [6]. 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 или переводами, соответствующими выбранным значениям.

Результаты систем сравниваются с базисными уровнями: поиск без расширений (noexp), и поиск с полным расширением – запросы расширяются синонимами, соответствующими всем возможным значениям (expall). В проведенных экспериментах в одноязычном поиске лучший результат был получен при поиске без расширения синонимами, в двуязычном информационном поиске использованием переводов по всем значениям expall.

Таким образом, в первом проведенном соревновании с использованием методов автоматического разрешения многозначности системам не удалось получить результаты, превышающие результаты методов, не использующих процедуру автоматического разрешения многозначности. Организаторы процедуры оценки связывают часть проблем с выбранной базовой системой поиска и намерены продолжать исследования роли автоматического разрешения многозначности в информационном поиске.

В работе [105] в качестве базовой модели информационного поиска используется формула Bm25, построенная на основе вероятностной модели информационного поиска (иначе называемая модель OKAPI [173]. Вес терма (слова) в документе вычисляется в этой модели по следующей формуле:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

где $f(q_i, D)$ – это частотность терма q_i в документе D, $|D|$ – это длина документа D в словах, *avgdl* – средняя длина документов в коллекции, k_1 и b – это параметры формулы, обычно принимающие значения k_1=2.0, b = 0.75. $IDF(q_i)$ (обратная частота встречаемости терма в документах коллекции) в данном случае вычисляется как:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

где N – это общее число документов в коллекции и $n(q_i)$ – число документов, содержащих q_i.

К модели OKAPI в работе [105] добавлен поиск по фразам и используется расширение запроса по отношениям WordNet. После разрешения многозначности слов к запросу добавляются синонимы, гипонимы и слова из определений синсетов. Основное свое внимание авторы концентрируют на коротких (двух или трехсловных) запросах.

Значение многозначного слова в запросе выбирается на основе толкований синсетов WordNet. При этом выбранные значения используются не для того, чтобы построить концептуальный индекс (индекс синсетов), а для того, чтобы найти подходящее расширение запроса.

Учитывая предшествующие неудачи использования WordNet для расширения запросов, авторы вводят дополнительные проверки возможности расширения, а также вес расширения. Важным элементом проверки возможности расширения запросов является предварительная оценка глобальной корреляции между отдельными словами.

Для оценки глобальной корреляции между словами используется следующая формула:

$$Global\_correlation\ (t_i, s) = idf\,(s) \cdot \log\,(dev\,(t_i, s))$$

$$dev\ (t_i,\ s) = \frac{co-occurrence\,(t_i, s) - df_i \cdot sdf / N}{df_i \cdot sdf / N}$$

где s – элемент запроса (отдельное слово или словарное выражение), t_i – некоторое другое выражение, df_i и sdf – это количество документов, содержащее t_i и s соответственно, N – число документов в коллекции, $idf\,(s)$ – обратная частота встречаемости s в коллекции, *co-occurrence* $(t_i,\ s)$ – число документов, в которых встречаются t_i и s, $dev\ (t_i,\ s)$ показывает степень отклонения совместной встречаемости t_i и s от независимого употребления.

Авторы предлагают расширять запрос, состоящий из двух термов t_1 и t_2, синонимами следующим образом.

Терм t_{11}, который является синонимом к терму запроса t_1 в синсете S, может быть добавлен в качестве расширения запроса, в одном из двух случаев:

- или S – является доминантным синсетом для терма t_{11}, т.е. t_{11} наиболее часто употребляется в значении, соответствующем синсету S;
- или t_2 имеет высокую степень корреляции с t_{11}, и величина корреляции между t_2 и t_{11} больше, чем величина корреляции между t_2 и t_1.
- При этом расширение производится со следующим весом:

$$w\ (t_{11}) = f\ (t_{11},\ S)/F(t_{11}) \qquad (*)$$

где $f\ (t_{11},\ S)$ – это частота встречаемости терма t_{11} в значении S, $F\ (t_{11})$ – это сумма всех частот для всех значений t_{11}. Частота значений берется из информации, приписанной синсетам в WordNet, которая, в свою очередь, получена на основе разметки текстового корпуса значениями WordNet. Этот вес интерпретируется как вероятность того, что терм t_{11} имеет значение S.

Для расширения запроса гипонимами проводятся проверки другого рода. Пусть U – синсет-гипоним для t_1. Синоним из U добавляется к запросу в следующих случаях:

1) U – это единственный гипоним синсета S терма t_1. Для каждого терма t_{11} из U этот терм добавляется к запросу, с весом (*), если U – это доминантный синсет t_{11};
2) U – это не единственный гипоним синсета S терма t_1, при этом определение U содержит либо термин t_2 или его синонимы. Тогда для каждого терма t_{11} из U этот терм добавляется к запросу, с весом (*), если U – это доминантный синсет t_{11}.

Авторы работы показывают на пяти разных текстовых коллекциях конференции TREC, что применение технологии разрешения многозначности к коротким запросам и на этой основе расширение запроса приводит к росту средней точности поиска от 4% до 34%.

Результаты по улучшению информационного поиска с использованием WordNet и информации о совместной встречаемости слов в рамках языковой (порождающей) модели информационного поиска получены в работе [27].

«Языковые порождающие модели» – это группа статистических методов, которые оценивают вероятность появления последовательности из m слов $P(w_1, \ldots, w_m)$ посредством вычисления вероятностного распределения.

В информационном поиске языковые модели используются для установления отношений между запросом Q и документами коллекции, в том смысле, что упорядочение документов при выдаче ответов на запрос определяется на основе оценки вероятности того, что языковая модель, построенная по документу, породит совокупность слов запроса $P(Q|M_d)$ [166, 189].

Основной формулой языковой модели информационного поиска для так называемой униграммной модели, т.е. в том случае, если все слова запроса рассматриваются как независимые друг от друга сущности, является следующая:

$$P(q_1, q_2, \ldots q_n \mid d) = \prod P(q_i|d)$$

Данная формула означает, что вероятность порождения запроса из документа в униграммной модели оценивается как произведение вероятности порождения отдельного элемента запроса из документа. Наиболее естественным способом оценки $P(q_i|d)$ является оценка вероятности встречаемости терма q_i в документе d посредством так называемой оценки максимального правдоподобия (maximal likelihood estimate – MLE), т.е.

$$P(q_i|d)=freq(q_i, d)/length(d)$$

Оценка вероятности последовательностей слов может оказаться достаточно сложной для текстовых коллекций, поскольку некоторые возможные последовательности слов могли никогда не встречаться в базовой коллекции, и не могли использоваться для качественной настройки языковой модели (training of language model), т.е. возникает - так называемая проблема нехватки данных (data sparceness). По этой причине важным элементом языковых моделей является процедура сглаживания (smoothing) [129, 224].

Большинство формул сглаживания предложено в рамках моделей, созданных для распознавания речи. В сфере языковых моделей для информационного поиска ситуация нехватки данных проявляется в том, что если элемент запроса не содержится в документе, то при выбранном способе оценки вероятности получается $P(q_i|d)=0$ и, следовательно,

$$P(q_1, q_2, ...q_n \mid d) = 0.$$

Одной из распространенных техник сглаживания является учет вероятности появления слова в коллекции $P(q_i|C)$, и тогда обобщенная формула сглаживания выглядит следующим образом:

$$P(w|d) = \begin{cases} P_s(w|d), & \text{если слово запроса встречалось в документе,} \\ \alpha_d P(w|C), & \text{если слово не встречалось в документе.} \end{cases}$$

где $P_s(w|d)$ – это сглаженная вероятность $P(w|d)$, $P(w|C)$ – это вероятность появления слова в коллекции, α_d – коэффициент учета каждой из моделей, в общем случае может зависеть от документа.

Один из простейших вариантов формулы, так называемое сглаживание Jelinek-Mercer, выглядит следующим образом:

$$P(q_i \mid M) = \lambda \cdot P(q_i \mid d) + (1 - \lambda) \cdot P(q_i \mid C)$$

Другим примером формулы сглаживания является так называемая формула абсолютного дисконтирования (absolute discounting). Идея метода заключается в понижении вероятности встреченных слов путем вычитания констант вместо умножения их на коэффициенты λ и (1-λ):

$$P(q_i \mid M) = \frac{\max(c(q_i, d) - \delta, 0))}{\sum_w c(w, d)} + \sigma \cdot P(q_i \mid C)$$

где M – модель сглаживания, δ – сглаживающая константа величиной от 0 до 1; $\sigma = \delta \cdot \frac{|d_u|}{|d|}$; $|d_u|$ – число уникальных слов в документе; |d| - общее количество слов в документе, т.е.

$$|d| = \sum_w c(w, d)$$

Учет $P(q_i|C)$ в языковых моделях играет роль, сходную с учетом обратной частотности (idf) в векторной модели информационного поиска [224]. Эксперименты в рамках конференции TREC [129, 166] показали эффективность языковых моделей для информационного поиска, однако существенным для эффективной работы методов является процедура подбора подходящей процедуры сглаживания. В работе [224] исследовались различные виды сглаживания. На основе этого исследования авторы делают выводы, что некоторые виды сглаживания в информационном поиске лучше подходят для коротких запросов, а другие для более длинных сложных запросов.

Авторы работы [27] подчеркивают, что классическая языковая модель информационного поиска основана на независимости слов в текстах друг от друга, что не соответствует реальному положению дел.

Информацию о взаимосвязи слов можно получить из двух источников:

- во-первых, подсчитывая совместную встречаемость слов в некотором текстовом окне.
- во-вторых, извлекая вручную описанные отношения из WordNet, поскольку некоторые указанные лингвистами отношения между словами может быть невозможно извлечь из рабочей коллекции. При этом отношениям из WordNet предлагается приписывать вес также на основе их совместной встречаемости в текстовом окне заданной величины.

Таким образом, оценивая вероятность порождения запроса из документа, предлагается использовать три источника информации по следующей формуле:

$$P(q|d)=\prod_{I=1} [\lambda_L P_L(qi|d) + \lambda_{CO} P_{CO}(qi|d) + \lambda_U P_U(qi|d)],$$

где $P_U(qi|d)$ – вероятность, полученная по классической униграммной языковой модели, – далее модель UM; $P_L(qi|d)$ – вероятность порождения запроса из документа, полученная на основе отношений лингвистического ресурса WordNet, – далее модель LM; $P_{CO}(qi|d)$ – вероятность порождения запроса из документа, полученная на основе совместной встречаемости двух слов в текстовом окне, – далее модель CM; λ_L, λ_{CO}, λ_U – подбираемые коэффициенты. В базовой униграммной языковой модели в качестве формулы сглаживания использовалась формула абсолютного дисконтирования.

Исследовался и другой вариант формулы, который приписывал отдельные веса разным типам связей WordNet: синонимам, гипонимам и гиперонимам:

$$P(q|d)=\prod_{I=1} [\lambda_1 P_{SYN}(qi|d) + \lambda_2 P_{HYPE}(qi|d) + \lambda_3 P_{HYPO}(qi|d) + \lambda_4 P_{CO}(qi|d) + \lambda_5 P_U(qi|d)],$$

где $\lambda_{1\ldots5}$ – весовые коэффициенты каждого типа отношений.

Совместная встречаемость слов, связанных между собой по WordNet, оценивалась в пределах абзаца. Совместная встречаемость слов, не поддержанных отношениями в WordNet, оценивалась в окне из 7 слов.

Для оценки совместной встречаемости в обоих случаях была также применена формула в духе языковых моделей с типом сглаживания по абсолютному дисконтированию. Так, формула для слов, между которыми описаны отношения в WordNet, такова:

$$P_L(w_i \mid w) = \frac{\max(c(w_i, w \mid W, L) - \delta, 0)}{\sum_{w_j} c(w_j, w \mid W, L)} + \frac{c(*, w \mid W, L)\delta}{\sum_{w_j} c(w_j, w \mid W, L)} P_{add-one}(w_i \mid W, L)$$

$$P_{add-one}(w_i \mid W, L) = \frac{\sum_{j=1}^{|v|} c(w_i, w_j \mid W, L) + 1}{\sum_{i=1}^{|v|} \sum_{j=1}^{|v|} (c(w_i, w_j \mid W, L) + 1)}$$

где $C\ (wi,w|W,L)$ – число совместных встречаемостей слов wi и w, связанных отношениями в WordNet, в пределах окна, $C(*, w|W,L)$ – число уникальных терминов встречающихся в окне W. Данная формула соответствует так называемой битермной языковой модели [193].

Предложенная модель тестировалась на текстовых коллекциях конференции по методам информационного поиска TREC общим размером более 1200 мегабайт. Оба варианта модели показали лучшие характеристики средней точности, по сравнению с базовой моделью. Большее увеличение показал второй вариант модели, который использовал разные весовые коэффициенты для разных типов отношений WordNet.

Анализ различных комбинаций подэлементов модели показал, что комбинация всех трех элементов модели (UM+LM+CM) всегда превышает показатели частичных комбинаций моделей. Это подтверждает мысль авторов, что посредством привлечения знаний из WordNet удалось использовать в поиске дополнительные сведения, которые не удалось получить на базе только использования информации о совместной встречаемости слов в текстовом окне.

В работе [181] исследуется вопрос, насколько велика точность работы систем информационного поиска в смысле обеспечения высокой точности выдачи в первых документах выдачи. Исследуя результаты поиска системы Lemur [157] по заголовкам запросов TREC, они показали, что только в 40% из 150 исследуемых запросов на первом месте поисковой выдачи находился релевантный документ.

Проанализировав причины такой ситуации, авторы работы установили, что для улучшения качества поиска необходимо производить расширение поискового запроса. При этом для обеспечения качественного расширения запроса необходимо определить, какие именно слова можно дополнить близкими по смыслу словами в контексте данного запроса, и какими именно из близких по смыслу слов. Так, включение в запрос многозначного слова может привести к резкому снижению качества поиска.

Для определения критериев расширения запроса близкими по смыслу словами авторы предлагают использовать показатель ясности (“clarity”) слов. Вычисление этого параметра основывается на следующих наблюдениях.

Если в ответ на запрос получены релевантные документы, то первые документы выдачи характеризуются относительно высокой частотностью небольшого числа тематических терминов. С другой стороны, если в ответ на запрос выдаются нерелевантные документы разнообразной тематики, то по распределению частот документы выдачи должны быть сходны с коллекцией в целом. В результате экспериментов было получено, что при поиске по заголовкам запросов мера точности поиска Precision (1) повысилась на 16.40% с 40.67% до 46.67%, средняя точность выросла на 0.89%. При поиске по полю описание (description) запроса Precision(1) повысилась на 18.18% с 44.00% до 52.00%, средняя точность выросла на 11.45%.

Таким образом, выборочное расширение запроса синонимами из WordNet привело к значимому улучшению результата поиска как по критерию Precision(1), так и по показателю средней точности.

1.2.4. Лингвистические онтологии в вопросно-ответных системах

1.2.4.1. Методы применения лингвистических онтологий в вопросно-ответных системах

Одним из активно развивающихся направлений в сфере информационного поиска является разработка вопросно-ответных систем. От традиционных информационно-поисковых систем вопросно-ответные системы отличаются тем, что должны предоставить пользователю не набор документов, которые наиболее релевантны поставленному вопросу, но выдать фрагмент текста, содержащий точный ответ на заданный вопрос.

В 1999 году стало проводиться тестирование вопросно-ответных систем в рамках конференции по информационному поиску TREC [210], в которых системы должны были искать ответы на вопросы вида: *What is the brightest star visible from the Earth? (Какая звезда, видимая с Земли, является самой яркой?)*

Основными этапами поиска ответа на вопрос в современных вопросно-ответных системах являются следующие (см. рис. 1.1):

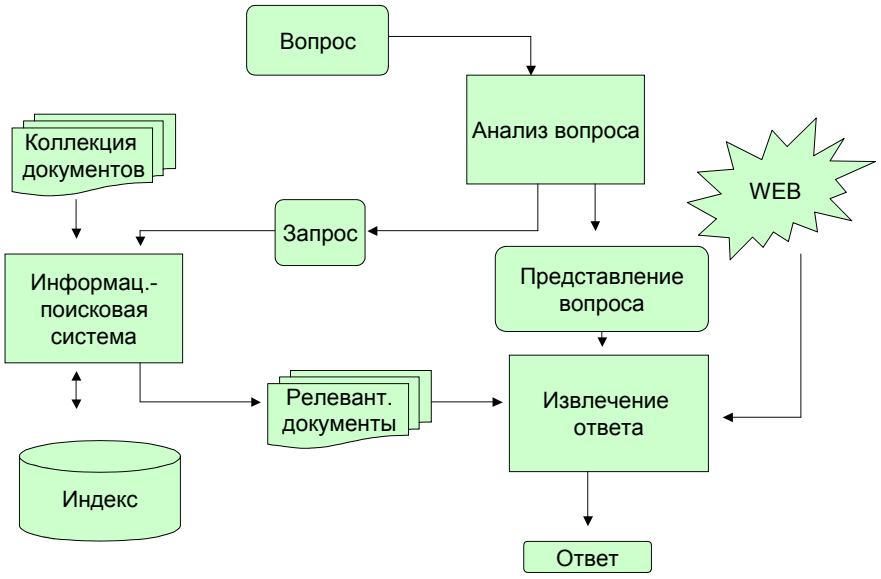


Рис. 1.1. Основные этапы обработки вопроса и формирования ответа вопросно-ответной системой

Прежде всего, производится подробный анализ вопроса, в результате которого определяется тип вопроса (вопрос времени, места, количества и другие) и соответствующий тип ответа, а также формируется запрос к информационно-поисковой системе. На втором этапе производится поиск релевантных документов или абзацев информационно-поисковой системой, формируется упорядоченный список наиболее релевантных документов (абзацев), из которого выбирается первых n (например, n=100-1000) документов (абзацев) для дальнейшей обработки. На третьем этапе производится подробный анализ полученных абзацев: содержит ли абзац требуемый тип ответа, близость слов ответа и вопроса, сходство синтаксических структур и т.п. В ходе такого анализа полученные абзацы оцениваются по мере возможности вхождения в них ответа на заданный вопрос, и переупорядочиваются на основе полученных оценок.

Обработка поискового запроса в рамках вопросно-ответной системы имеет свою специфику по сравнению с обработкой типичного запроса при поиске в Интернет. Как известно, запросы в глобальных информационно-поисковых системах обычно очень короткие – 2-3 слова, и по ним находятся сотни и тысячи документов. Запросы в форме вопросов обычно значительно длиннее, поэтому если требовать присутствия в документе сразу всех слов запроса, то чаще всего не будет найдено ни одного документа, что означает, что поисковая система должна автоматически определить, какие слова такого запроса должны быть отброшены или заменены.

Классическая векторная модель на основе сравнения векторов запроса и документа позволяет найти наиболее релевантные документы и по частично совпадающему запросу [306]. Однако при формальном выполнении пословных векторных моделей важные для ответа слова вопроса могут быть автоматически отброшены, поэтому в некоторых современных исследованиях по вопросно-ответным системам стали использоваться не векторные модели поиска, а выполняется булевский поиск.

В булевской модели поиска слова запроса соединяются между собой логическими связками: *AND* (&), *OR*(∨), *NOT*(¬), которые могут быть сгруппированы при помощи скобок. Таким образом, запрос пользователя представляется логической формулой, в которой атомами могут быть термины или какие-либо дополнительные условия (например, тип коллекции или документа, ограничение на расстояние между словами запроса и т.п.). Каждый атом формулы соответствует булевской функции, проверяющей вхождение заданного термина или выполнение заданного условия в анализируемом документе.

Поисковая машина, основанная на булевом поиске, возвращает документы, для которых формула запроса принимает истинные значения. Каждому атому формулы сопоставляется множество документов, для которых значение атома истинно. Если атом является термином, то ему сопоставляется множество документов, в которых термин встречается. Затем над множествами выполняются элементарные операции — объединения, пересечения и дополнения, соответствующие логическим связкам между атомами.

Использование булевской модели поиска, которая при выполнении стандартного информационного поиска, считается менее качественной, чем векторная модель [130], связано с тем, что при выполнении задачи сокращения формулировки запроса необходимо осуществлять дополнительный контроль, какие слова формулировки вопроса обязательно должны присутствовать в тексте ответа, а какие могут быть пропасть в тексте ответа с минимальным ущербом для релевантности ответа [71, 82, 93].

Булевский запрос обычно формируется как конъюнкция всех значимых слов формулировки вопроса. Если проводится морфологический анализ запроса или добавляются синонимы, то они объединяются в дизъюнкцию. Поскольку стандартной является ситуация, когда не находится документов, которые содержат все значимые слова вопроса, поэтому при обработке вопроса часто необходимо определить, какие именно слова формулировки

вопроса можно отбросить, не включить в поисковый запрос без потери сути вопроса.

Так, упомянутый в начале раздела вопрос «Какая звезда, видимая с Земли, является самой яркой» может быть преобразован в следующий булевский запрос, в котором часть слов из запроса пропадает, а близкие по смыслу слова образуют дизъюнкции:

ЗВЕЗДА AND (ЯРКИЙ OR ЯРЧАЙШИЙ) AND ЗЕМЛЯ

Для более точного определения, какие именно слова могут формулировки вопроса могут быть отброшены, обычно предлагается система модификаций, упрощающих исходный булевский запрос, после каждой из которых опять происходит обращение к поисковой системе для проверки, не появились ли релевантные документы.

Обычно используются два основных способа упрощения булевского запроса. Во-первых, можно часть конъюнкций переводить в дизъюнкции. Вторым способом является поочередное исключение членов конъюнкции, на основе некоторого множества эвристик, определяющих значимость членов конъюнкции.

В связи с длинной формулировкой естественно-языкового вопроса и частым отсутствием в самых больших текстовых коллекциях ответов, содержащих все или большинство слов формулировки вопроса, значимой становится роль лексических ресурсов, позволяющих найти ответы в тех предложениях, в которых часть слов заменена на близкие по смыслу слова. Таким образом, роль лексических ресурсов, онтологий, тезаурусов при обработке вопросов в вопросно-ответных системах представляется достаточно важной.

Многие современные вопросно-ответные системы используют в качестве лексического источника WordNet. В таких системах WordNet может использоваться для решения следующих задач:

- распознавания типа вопроса;

- классификации типов ответов;
- для реализации лексических и семантических замен.

Одной из самых эффективных систем в вопросно-ответной дорожке конференции TREC 1999 стала вопросно-ответная система Южного Методистского университета, которая на нескольких этапах обработки вопроса и поиска ответа обращается к информации, хранимой в тезаурусе WordNet [71].

Лексические и семантические замены осуществляются в момент сопоставления формальной структуры вопроса и ответа. Поиск в системе организован на основе обработки булевских запросов, в качестве единиц поиска выступают не целые документы, а абзацы [71].

На этапе обработки вопроса WordNet используется для определения типа вопроса и типа ответа. Например, если вопрос начинается со слов «*what company*» – этот вопрос классифицируется как вопрос об организации.

На некоторые типы вопросов, кандидаты-ответы могут быть получены непосредственно из WordNet. Например, если задан такой вопрос как «*What flowers did Van Gogh paint?*» (*Какие цветы рисовал Ван Гог),* то может быть извлечен список всех 470 видов цветов, упомянутых в WordNet, и использован для проверки в качестве подходящего ответа.

Для организации поиска ответов была разработана классификация ответов на вопросы конференции TREC, которая включала такие типы, как: *время, дата, продукция, организация, деньги, место, язык, человек.*

После этого WordNet был преобразован в таксономию ответов, релевантные синсеты были сгруппированы под своим типом ответа, а нерелевантные синсеты были удалены. В результате полученная таксономия ответов включала 8707 синсетов, 20 верхних типов. Было добавлено 129 отношений, отсутствующих в WordNet, но полезных для ответов на вопрос.

Таким образом, в значительной мере для нужд классификации вопросов и ответов на основе информации WordNet был построен новый ресурс, настроенный на вопросы, предлагаемые в рамках конференции

TREC. На основе проделанной работы была достигнута правильная идентификация типа ответа для 79% вопросов на конференции TREC-9.

В данной вопросно-ответной системе тезаурус WordNet совместно с серией булевских запросов используется для подбора необходимых лексических и семантических замен. При обработке формулировки запроса строится синтаксическая структура предложения, которая называется семантической формой запроса, а также создается булевское выражение, состоящее из слов запроса. Выполняется поиск, и отбираются абзацы текста, удовлетворяющие запросу и содержащие, по крайней мере, одно языковое выражение, подходящее по типу к требуемому типу ответа. После этого могут быть инициализированы три цикла расширения запроса, в процессе которого булевское выражение запроса пополняется семантически близкими словами из WordNet, которые связаны между собой в запросе дизъюнктивно.

В результате, качество поиска ответов на вопросы TREC при возвращении короткого 50-байтного ответа улучшилось на 76%.

1.2.4.2. Предметные области вопросно-ответных систем

Современные вопросно-ответные системы можно подразделить на два больших класса.

Первый класс – это вопросно-ответные системы общего назначения, которые должны отвечать на широкий круг вопросов на базе сверхбольших текстовых коллекций, например, информации, хранящейся на интернет-сайтах. Величина используемых текстовых коллекций часто позволяет такой системе воспользоваться избыточностью информации, и находить такой текст, в котором ответ может быть получен системой наилучшим образом.

Второй класс вопросно-ответных систем – это вопросно-ответные системы, созданные для ответов на вопросы в рамках конкретных предметных областей, например, поиска информации в технической документации, в коллекции ответов на частые вопросы пользователей и другие. Такие системы располагают значительно меньшей коллекцией

документов. В значительной мере для качественного поиска ответов на вопросы эти системы должны пользоваться знаниями о предметной области, хранимых, в частности, в форме онтологий и тезаурусов [143].

Примерами сфер приложений специальных вопросно-ответных систем являются правовая сфера, а также многочисленные форумы по техническим проблемам, программному обеспечению, куда обращаются пользователи со своими проблемами.

Представляется, что сужение сферы деятельности позволяет точнее настроить вопросно-ответную систему, и это действительно так. Однако в предметных областях возникает другая проблема: реальные вопросы пользователей не представляют собой аккуратно построенный в виде одного предложения вопрос. Чаще, вопрос реального пользователя включает предварительное описание проблемной ситуации, своих действий в этой ситуации, может содержать несколько подвопросов с отдельными вопросительными словами, а также может содержать значительно количество вводных слов, и другого рода, бессодержательных слов (*помогите, пожалуйста, поясните, help* и т.п.).

Приведем пример реального вопроса в правовой области: •

Расскажите, пожалуйста, о туристических и транзитных визах в США. Что собой представляют визы, выдаваемые супругам, и визы, связанные с обучением? Сколько стоит оформление визы?

В работе [86] указывается, что «если современные интернет-поисковики демонстрируют достаточно высокое качество обработки 2-3 словных запросов, их способность отвечать на сложные вопросы... является явно недостаточными».

Авторы работы [103] также пишут о том, что исследования вопросно-ответных систем в рамках TREC в наибольшей степени было сконцентрировано на коротких, направленных на поиск фактов общезначимых вопросов, поиск ответов на многие из которых базируется на

избытке информации в интернет. Предложенные подходы достаточно хорошо работают для вопросов типа TREC, однако хорошие результаты не обязательно обеспечивают успех при обработке вопросов вне конференции TREC**.**

В [103] описывается система обработки реальных вопросов в рамках более широкой области аэрокосмической индустрии. Основные компоненты вопросно-ответной системы включают: 1)обработка документов 2) модуль язык – логика (L2L) 3) поисковая машина и 4) нахождение абзацев с ответом. Когда пользователь спрашивает систему, его вопрос сначала посылается в L2L модуль, который порождает внутреннее представление вопроса и идентифицирует фокус вопроса. Поисковая машина возвращает 50 лучших документов. В качестве ответов возвращается 20 лучших абзацев.

Вопросы NASA отличаются от вопросов TREC в нескольких аспектах. Во-первых, вопросы NASA задаются в реальное время студентом, и вопрос может быть многозначным или предполагает неявное знание, которое не эксплицировано в вопросе. Реальные вопросы обычно пишутся в спешке и могут быть сформулированы с нарушением грамматической структуры или содержать орфографические ошибки.

1.2.4.3. Поиск ответов на вопрос в вопросно-ответных сервисах

Отдельным направлением в развитии вопросно-ответных систем может рассматриваться поиск уже существующих ответов в вопросо-ответных сервисах глобальных интернет-поисковиков.

Во многих странах стали популярными вопросно-ответные сервисы, когда пользователь может обратиться к сообществу пользователей или к экспертам за ответом на свой вопрос. Такие службы обычно накапливают большие объемы уже отвеченных вопросов, т.е. документов типа «вопрос-ответ». При задании вопроса сервис может, прежде всего, выполнить поиск на предмет того, нет ли уже в его базе вопросно-ответных документов ответа на подобный вопрос.

Вместе с тем такие вопросы, будучи сходными по значению, могут быть сформулированы с помощью совершенно разных лексических средств. [86] приводят такие примеры близких по содержанию вопросов, не содержащих ни одного общего слова:

1. Is downloading movies illegal?

2. Can I share a copy of a DVD online?

Поиск ответов на такие вопросы отличается от основной парадигмы современных вопросно-ответных систем тем, что нужно найти не короткий ответ на относительно ограниченный список типов вопросов, а документ, отвечающий на неограниченный список типов вопросов.

1.2.5. Лингвистические онтологии в системах автоматической рубрикации текстов

Классификация/рубрикация информации (отнесение порции информации к одной или нескольким категориям из ограниченного множества) является традиционной задачей организации знаний и обмена информацией, рассматривается как одна из классических задач информационного поиска. Распространенность больших информационных коллекций делает необходимым развитие автоматических методов рубрикации.

Таким образом, имеется следующая постановка задачи [180]:

- имеется множество категорий (классов, меток): $C = \{c_1,...c_{|c|}\}$,
- имеется множество документов: $D = \{d_1,...d_{|d|}\}$,
- существует неизвестная целевая функция Ф: $C \times D \rightarrow \{0, 1\}$

Необходимо построить классификатор Ф', максимально близкий к Ф.

1.2.5.1. Методы автоматической рубрикации

Известны две основных технологии автоматической рубрикации:

- методы, основанные на знаниях (также именуемые "инженерный подход"), при применении которых правила отнесения текстов к рубрикам строятся инженерами по знаниям в форме булевских выражений, правил продукций и т.п.
- методы на основе машинного обучения, при применении которых используется коллекция документов, предварительно отрубрицированная человеком.

Алгоритм машинного обучения строит процедуру классификации документов на основе автоматического анализа заданного множества отрубрицированных текстов, т.е. имеется некоторая начальная коллекция размеченных документов $R \subset C \times D$, для которых известны значения Ф. Данная размеченная коллекция делится на обучающую и тестировочную части. Первая используется для обучения классификатора, вторая – для проверки качества классификации.

Классификаторы обоих типов могут выдавать точный ответ Φ': $C \times D \rightarrow \{0, 1\}$ или степень подобия $\Phi': C \times D \rightarrow [0, 1]$ [180].

Оценка качества автоматической классификации производится путем сравнения с эталонной («правильной») классификацией набора документов, т.е. на основе коллекции документов, отрубрицированных вручную. Для оценки эффективности работы систем рубрицирования используются такие характеристики, как полнота, точность, F-мера, аккуратность [227].

Полнота (r – recall) - это отношение R/Q, где R - количество текстов, правильно отнесенных к некоторой рубрике, а Q – общее количество текстов, которые должны быть отнесены к этой рубрике.

Точность (p – precision) – это отношение R/L, где R – количество текстов, правильно отнесенных системой к некоторой рубрике, а L – общее количество текстов, отнесенных системой к этой рубрике.

Метрика F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера для данного запроса (рубрики) вычисляется по формуле:

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Также иногда используется метрика аккуратности (accuracy), которая вычисляется как отношение правильно принятых системой решений к общему числу решений. Формально

$$Аккуратность = (R+R-)/D,$$

где R – количество текстов, правильно отнесенных системой к рубрике, R – число текстов, правильно не отнесенных системой к рубрике, D – общее число документов в коллекции. Таким образом, знаменатель не зависит от рассматриваемой рубрики.

Считается, что наиболее эффективными, но и наиболее трудозатратными, является методы автоматического рубрицирования, основанные на знаниях. При рубрицировании текстов на основе знаний используются заранее сформированные базы знаний, в которых описываются языковые выражения, соответствующие той или иной рубрике, правила выбора между рубриками и др. [74].

Так, например, в классической работе по инженерному подходу к автоматической рубрикации текстов [74] рубрики определяются на основе сопоставления каждой рубрике совокупности специальных шаблонов. Шаблон определяется как конструкция, состоящая из произвольного количества дизъюнкций, конъюнкций, отрицаний, пропусков слов и операторов необязательности. В такой конструкции могут быть также заданы части речи, способ написания (с большой или маленькой буквы), знаки препинания. Каждому такому шаблону приписан вес, определяющий, насколько сильно этот шаблон соответствует той или иной рубрике. Суммирование весов шаблонов, сопоставленных одной и той же рубрике по тексту, дает величину соответствия этой рубрики тексту. Решение о выборе рубрик для текста принимаются на основе правил, в которых учитывается,

какие рубрики были обнаружены в тексте, в какой части текста встречались соответствующие шаблоны, и какой суммарный вес имеет каждая рубрика.

Результаты работы таких систем на тех текстовых потоках, для которых они проектировались, дают очень высокие оценки эффективности автоматического рубрицирования. Например, в работе [74] приводятся следующие характеристики эффективности работы системы автоматического рубрицирования экономических и финансовых сообщений информационного агентства Рейтер: точность – 84%, полнота – 94%. Объем рубрикатора – 674 рубрики. В работе [172] сообщается о реализации технологии автоматической рубрикации, достигающей 100% точности при 60% полноты.

Однако разработка систем автоматического рубрицирования, основанных на знаниях, требует больших затрат труда и часто занимает несколько человеко-лет. В таких системах базы знаний и алгоритмы жестко настроены не только на предметную область, но и на рубрикатор, размер и формат текстов. Поэтому изменение рубрикатора или необходимость рубрицирования текстов той же предметной области, но из другого источника информации влечет за собой значительные дополнительные усилия.

В настоящее время можно наблюдать всплеск научных работ, посвященных применению методов машинного обучения для автоматической рубрикации текстов. Приводятся высокие оценки результатов работы таких методов [47, 88, 99, 222]. Однако, как отмечалось в ряде работ [2, 174, 215], для больших рубрикаторов – 500 и более тематических рубрик – из-за трудности формирования качественной непротиворечивой обучающей коллекции во многих случаях работающим подходом в настоящее время является так называемый подход, основанный на знаниях [74, 215, 251], подразумевающий ручное описание смысла каждой рубрики. Например, в компании Рейтер, предоставляющей текстовые коллекции, на которых продемонстрированы многие высокие результаты технологий машинного обучения, в собственном бизнес-процессе

используется технология, сочетающая работу системы автоматической рубрикации, основанной на знаниях, с последующим просмотром редакторами [174].

Проблемы машинного обучения для рубрикации текстов связаны, в частности, с тем, что при разработке таких систем необходима коллекция документов, размеченная экспертами по рубрикам. Для эффективного обучения рубрицированию по большому рубрикатору требуется большее число размеченных документов. Важной особенностью такой размеченной коллекции является то, что разметка должна быть выполнена последовательно, т.е. необходимо, чтобы эксперты применяли одни и те же принципы отнесения текстов к рубрике, чтобы похожие документы получали похожие рубрики.

1.2.5.2. Использование знаний в автоматической рубрикации текстов

Подходы машинного обучения для автоматической рубрикации документов используют для своего обучения набор свойств, характеристик исходного документа. Существенной составной частью этих свойств является множество слов (отличных от стоп-слов), упоминаемых в документах. Одним из направлений в подходах, стремящихся увеличить предсказуемостную мощность обучающего метода, является использование знаний о синонимах и лексических отношениях, описанных в WordNet.

Наиболее популярным направлением исследований привлечения информации из WordNet для автоматической рубрикации текстов является дополнение пословного представления документа в виде векторной модели синсетами из WordNet, после чего применяется тот или иной метод машинного обучения.

Одной из первых работ, в которой авторы пытались интегрировать лексическую информацию из WordNet в набор характеристик для машинного обучения, была работа [23]. В этой работе было выдвинуто предположение,

что обучаемая модель может быть усилена за счет применения синонимов к заголовкам категорий, используемых для рубрикации. Для этого авторы вручную выбрали подходящие синсеты из WordNet. Применялось два метода машинного обучения: метод Rocchio и метод Widrow-Hoff. Сравнение этих методов, обученных только на векторах слов, и с учетом названий рубрик и их синонимов, проводилось на коллекции Reuters-21578. Для обоих методов интегрированное представление дало значимое улучшение, особенно значительным улучшение было на рубриках с малым числом обучающих примеров (<10).

В работе [179] WordNet используется для расширения представления документа на базе всех слов документа. Разрешение лексической многозначности не производится, а берутся все синсеты слов, встретившихся в документе. Кроме того, вектор синсетов дополняется гиперонимами. Это дополнение регулируется параметром h – числом шагов обобщения. Использовался алгоритм обучения Ripper. Тестирование на нескольких коллекциях показало, что ни вектор из синсетов (h=0), ни вектор с одним уровнем обобщения не дали стабильного улучшения на разных коллекциях.

В работе [85] также используются синсеты и гиперонимы, но из всех синсетов многозначного слова выбирается наиболее частотный по коллекции синсет и соответствующий ему гипероним. Три алгоритма машинного обучения использовались для классификации текстов на базе различных комбинаций характеристик: слов, синсетов, синсетов с гиперонимами, биграмм. Эксперименты проводились на трех разных коллекциях. Авторы делают вывод, что использование гиперонимов привело к улучшению показателей автоматической рубрикации на всех коллекциях, и, кроме того, использование гиперонимов всегда улучшает показатели по сравнению с применением только исходных синсетов.

В работе [89] сравнивается качество автоматической рубрикации трех алгоритмов машинного обучения, включая Naïve Bayes и k-NN классификаторы на Брауновском корпусе, который размечен значениями

WordNet. Тексты корпуса разделены на 15 категорий, и собственно, этой классификацию и должны осуществлять классификаторы. Было отмечено, что результаты всех методов улучшились на множестве синсетов по сравнению с пословной базой обучения, однако это улучшение было слишком незначительным.

Влияние трех разных онтологических ресурсов на качество автоматической рубрикации изучалось в работе [81]. Исследовались такие ресурсы, как WordNet, онтология тезауруса в медицинской области MESH (22 тысячи понятий с синонимами и квазисинонимами) и тезаурус по сельскохозяйственной тематике AGROVOC (17 тысяч понятий) [7]. Исследование проводилось на базе метода машинного обучения AdaBoost.

Эксперименты на коллекции Reuters для 50 рубрик с наибольшим числом положительных примеров проводились с использованием синсетов и гиперонимов WordNet. На комбинированном представлении слова+синсеты+гиперонимы (5 уровней) было получено улучшение меры F1 на 3.29% (макроусреднение) и 2% (микроусреднение), что означает, что увеличение качества рубрикации было больше для рубрик с небольшим числом положительных примеров.

Медицинская онтология применялась для классификации текстов из коллекции OHSUMED. Здесь также использовались 50 рубрик с наибольшим числом примеров. Для обработки этой коллекции использовался также и WordNet. Разные варианты применения WordNet дали увеличение F1 меры от 2 до 7%. Относительное увеличение F1 меры на основе медицинской онтологии дало 3-5% на разных прогонах. Также увеличение F1 меры было достигнуто на некоторых прогонах для текстов сельскохозяйственной тематики на базе тезауруса AGROVOC (до 10% F1 меры).

В работе [130] исследуется влияние различных типов расширения по отношениям WordNet в задаче отнесения множества документов к одной из двух рубрик. 15 пар рубрик взято из нескольких коллекций, используемых для оценки качества автоматической рубрикации: Reuters-21578, USENET,

DigiTrad, Newsgroups. Для экспериментов использовались два классификатора: Naïve Bayes и SVM. Были сделаны отдельные прогоны для базовой пословной модели, расширения синонимами, расширения синонимами и гиперонимами, синонимами и гипонимами, синонимами и меронимами, синонимами и холонимами. Все расширения проводились только для существительных. В случае многозначных слов бралось наиболее частотное значение.

Авторы работы делают вывод, что расширение на гипонимы и меронимы (части) дает устойчивое снижение показателя «аккуратности» (accuracy), все остальные расширения не показывают значимого повышения показателя по сравнению с базовым классификатором.

Таким образом, на текущий момент разные исследования расходятся в мнениях по поводу того, насколько WordNet и другие онтологические ресурсы могут улучшить качество автоматической рубрикации при использовании их в качестве источник дополнительных знаний для машинного обучения. Некоторые работы показывают небольшое улучшение качества рубрикации, другие – не выявили никакого улучшения качества или неустойчивое улучшение.

Заключение к главе 1

Проведенный обзор применения тезаурусов в задачах информационного поиска показывает технологическую сложность этой проблемы. Так при появлении в открытом доступе в сети Интернет тезауруса WordNet многие исследователи предположили, что использование этого ресурса непременно должно улучшать качество информационного поиска, поскольку WordNet предоставляет большое количество дополнительной информации о словах, их синонимах, значениях, отношениях.

Однако многочисленные первые эксперименты по интеграции WordNet в информационный поиск закончились неудачей. Понадобилось практически 10 лет, чтобы предложить модели, в которых применение WordNet дало

значимое улучшение качества информационного поиска. Основной смысл предложенных удачных моделей заключается в том, что информация, полученная из WordNet, должна дополнительно взвешиваться, дополнительно оцениваются на основе особенностей конкретной коллекции, на которой производится поиск. Таким образом, производится как бы настройка WordNet на конкретную коллекцию и типовые запросы к этой коллекции.

Глава 2. Модель лингвистической онтологии для автоматической обработки текстов

2.1. Основные принципы разработки лингвистических ресурсов для автоматической обработки текстов

В предыдущей главе было показано, что для приложений информационного поиска использовались разные лингвистические и онтологические ресурсы: информационно-поисковые тезаурусы, тезаурусы типа WordNet, формальные онтологии. Все из них имеют некоторые проблемы при использовании их как ресурсов в рамках решения задач информационного поиска.

Традиционные информационно-поисковые тезаурусы создавались как инструмент для помощи человеку, их структура направлена на предоставление удобств индексатору (удаление слишком конкретных терминов, удаление близких по смыслу терминов, добавление комментариев по употреблению тех или иных дескрипторов). В связи с этим при использовании традиционных информационно-поисковых тезаурусов в автоматической обработке текстовой информации возникают существенные проблемы. В литературе предлагается использовать методы машинного обучения для проставления дескрипторов тезауруса по уже проиндексированному людьми множеству документов, создание которого представляется чрезвычайно дорогой процедурой.

Формальные онтологии, одним из провозглашаемых принципов которых является независимость от конкретного языка, сложно использовать в автоматической обработке текстов для приложений информационного поиска, поскольку для этого единицы формальной онтологии необходимо связать с единицами конкретного естественного языка [159, 162]. Кроме того, стремление к четкой формализации отношений между понятиями к формальной онтологии чрезвычайно трудно соблюсти в ситуации, когда

необходимо создавать сверхбольшие ресурсы, и, кроме того, приводит к проблемам при установлении связей «понятие – языковое выражение».

Ресурсы типа WordNet создаются для описания лексики языка в соответствии с лингвистическими традициями [97, 138]. Но любая информационная система имеет дела не только с общей лексикой, но и с конкретными предметными областями и их терминологиями. Анализируя попытки создать терминологические ресурсы на основе WordNet, следует отметить, что структура WordNet не приспособлена для описания терминологий. Раздельное описание частей речи, слишком большой набор несвязанных между собой значений, недостаточная проработанность принципов включения многословных выражений, – все это приводит к проблемам разработки и использования терминологических ресурсов, созданных на базе модели WordNet.

Вместе с тем, в каждом из этих типов ресурсов есть те качества, которые должны присутствовать в большом лингвистическом ресурсе для информационно-поисковых приложений. Такой ресурс для автоматической обработки текстов в информационно-поисковых приложениях в широких предметных областях должен сочетать принципы различных традиций и методологий:

- методологии разработки традиционных информационно-поисковых тезаурусов;
- методологии разработки лингвистических ресурсов типа WordNet (Принстонский университет);
- методологии созданий формальных онтологий.

Поскольку важно уметь описывать терминологию широких предметных областей, то необходимо использовать опыт разработки информационно-поисковых тезаурусов, а именно:

- информационно-поисковый контекст;
- единицы ресурса создаются на основе значений терминов;

- описание большого числа многословных выражений, принципы включения (невключения) многословных единиц;
- небольшой набор отношений между понятийными единицами.

Так как предполагается использовать лингвистический ресурс в автоматическом режиме обработки текстов, то необходимо использовать методологию разработки лексических ресурсов типа WordNet, в которой важны следующие положения:

- понятийные единицы создаются на основе значений реально существующих языковых выражений;
- многоступенчатое иерархическое построение лексико-терминологической системы понятий;
- принципы описания значений многозначных слов и выражений.

Из методологии разработки формальных онтологий важны следующие положения:

- разработка лингвистической онтологии как иерархической системы понятий [18, 132, 185];
- использование для описания отношений формально определяемых отношений с формальными свойствами;
- в качестве аксиом (правил вывода) использование свойств транзитивности и наследования отношений между понятиями.

Таким образом, в результате исследований и экспериментов сформулированы следующие принципы создания онтологических ресурсов для автоматической обработки текстов (далее *ЛО – лингвистическая онтология для автоматической обработки текстов)*.

1) Онтологию ЛО для предметной области *D* можно формально представить следующим образом:

$$ЛО=<C, Ex, NO, R_{lo}, A_{tr,i}, S, T, M_{m,a}, L, DC>$$

где C – множество понятий онтологии, где понятие обозначает класс сущностей, обладающих одинаковыми свойствами и отношениями с другими классами сущностей;

Ex – множество экземпляров понятий онтологии, задано отображение $E: C \rightarrow 2^{Ex}$;

NO – множество имен понятий и экземпляров в онтологии, имена уникальны;

R_{lo} – набор отношений между понятиями $R \subset C$ x C, специально разработанный для автоматической обработки текстов в широких предметных областях;

$A_{tr,i}$ – множество правил вывода, основанных на свойствах транзитивности и наследования отношений;

T – множество текстовых входов онтологии – языковых выражений, значения которых представлены в онтологии;

S – множество отношений между языковыми выражениями (T) и понятиями (C): $\{s\ (c_i, t_j)\}$;

$M_{m,a}$ – множество многозначных слов и выражений из T: $M_{m,a} \subset T$; многозначные текстовые входы онтологии делятся на два подвида: M_m – текстовые входы, которые относятся к более одному понятию онтологии, и M_a – текстовые входы, которые многозначны, но в онтологии представлено только одно значение: $M_{m,a} = M_m \cup M_a$

L – множество лемматических представлений языкового выражения (т.е. представление выражения в виде последовательности слов в словарной форме, например, словосочетания *ценная бумага* представляется в лемматическом виде как *ЦЕННЫЙ БУМАГА*),

DC – это отображение терминологического состава (TD) заданной коллекции предметной области $(Dcoll)$ на текстовые входы и понятия онтологии:

$$DC: (Dcoll, TD) \rightarrow (T, C).$$

Отображение DC задает критерий минимальной полноты онтологии, которая должна обеспечивать покрытие терминологического состава

заданной коллекции предметной области, что собственно и отражает суть лингвистической онтологии.

Рассмотрим компоненты модели подробнее.

Предлагаемая модель лингвистической онтологии близка к модели WordNet тем, что провозглашается необходимость подробного покрытия представленных в текстах предметной области понятий и способов их лексико-терминологического выражения в тексте. Отличие от модели WordNet заключается в том, что снижается зависимость создаваемой системы понятий от собственно языковых факторов таких, как разделение системы понятий по частям речи и синонимическая эквивалентность при замене в различных контекстах как фактор выделения понятий.

Таким образом, *лингвистическая онтология предметной области представляет собой базу знаний онтологического типа о понятийной системе и лексико-терминологическом составе предметной области.*

2) Единицей онтологии ЛО является понятие, как единица в системе понятий, имеющая свои специфические свойства, отличающие данную единицу от других единиц в системе понятий. Такой взгляд соответствует как современной трактовке дескрипторов в информационно-поисковых тезаурусах, так и понятий (классов) в онтологиях [117, 258, 299].

3) Каждое введенное понятие должно иметь однозначное имя. Именем может являться однозначное слово или словосочетание, значение которого соответствует этому понятию (т.е. один из текстовых входов понятия). Кроме того, имя может формироваться из многозначного текстового входа с сужающей значение пометой, или совокупностью синонимов, которая определяет значение однозначно. Подобная практика однозначного называния дескриптора подробно разработана в стандартах по созданию информационно-поисковых тезаурусов. Поскольку имя фиксирует особенности обозначаемого им понятия, то это соответствует и практике разработки формальных онтологий [299].

4) Каждое понятие снабжается набором текстовых входов – языковых выражений, значения которых соответствуют данному понятию. Такие языковые выражения являются между собой онтологическими синонимами. В текстах может встречаться множество вариантов текстовых входов того или иного понятия, как, например, известно о существовании множественной вариативности терминов предметной области. Эти варианты необходимо фиксировать сразу при вводе понятия, или дополнять при обнаружении в конкретном тексте, поскольку известно, что автоматические методы сопоставления терминов с учетом потенциальной вариативности приводят к снижению точности сопоставления [263].

5) В текстах предметной области значительную часть представляют слова, которые не принадлежат конкретной предметной области, могут встречаться в текстах многих предметных областях, т.е. принадлежащие общему лексикону *GL*, например, *создавать, участвовать, принимать* и многие другие. Поэтому многозначные слова, описанные в ЛО, делятся на две множества. В первое множество M_m входят выражения, которые отнесены более чем к двум понятиям в ЛО, например, *дерево как растение и дерево как материал*. Во второе множество M_a входят выражения, которые отнесены к одному понятию из ЛО, но данные слова могут иметь другое значение в *GL* (например, *стали: сталь, стать*), что отмечается специальной пометкой многозначности. Таким образом,

$$M = M_m \cup M_a$$

$$\forall t_i((t_i \in M_m) \rightarrow (\exists c_i, c_j \in C^D,\ c_i \neq c_j,\ (s(c_i, t_i) \wedge s(c_j, t_j))))$$

$$\forall t_i((t_i \in M_a) \rightarrow (\exists c_i \in C^D,\ c_j \in C^{GL}\ (s(c_{i,} t_i) \wedge s(c_j, t_j)))),$$

где C^{GL} – система понятий общего лексикона, которые, возможно, не описаны в данной ЛО, C^D – система понятий предметной области, для которой создается онтология, $s(c_{i,} t_i)$ – пара текстовое выражение и понятие онтологии, соответствующее значению этого текстового выражения.

6) Система отношений, используемых в лингвистической онтологии, представляет собой небольшой набор отношений, и в этом предлагаемая модель лингвистической онтологии близка к традиционным информационно-поисковым тезаурусам. Однако для установления отношений применяются строгие онтологические критерии. С каждым отношением связан свой набор аксиом (правил вывода), которые имеют важное значение для различных этапов автоматической обработки текстов и приложений информационного поиска [290].

Отношения между понятиями, описываемые в онтологическом ресурсе, предназначенном для автоматической обработки текстов в рамках информационно-поисковых приложений должны выполнять разнообразные функции.

Во-первых, эти отношения должны использоваться в классических функциях информационно-поисковых тезаурусов для расширения поискового запроса или вывода рубрики документа. Во-вторых, отношения важны для разрешения многозначности языковых единиц, включенных в ресурс, поскольку естественным методом реализации автоматической процедуры разрешения многозначности является сопоставление контекста употребления многозначной единицы в тексте и контекста соответствующего понятия в онтологическом ресурсе. В-третьих, отношения в онтологическом ресурсе могут использоваться для выявления лексической связности в текстах, и использованию выявленной структуры текста для улучшения качества обработки текстов.

Для реализации любой из этих функций необходимо осуществление специализированного логического вывода: встретив вхождение некоторого понятия в тексте, нужно делать многошаговые проходы по отношениям. В условиях широкой предметной области и, следовательно, необходимости создания лингвистической онтологии большой величины, для обработки текстов, не ограниченных по стилю, жанру, величине, наиболее стабильно можно опираться на те отношения, которые не исчезают, не меняются в

течение всего срока существования любого или подавляющего большинства экземпляров понятия, так лес всегда состоит из деревьев.

Поэтому в лингвистической онтологии описываются отношения только между такими понятиями c_i и c_j, которые присущи по крайней мере одному из этих понятий по определению (см. п. 2.2)

7) В качестве аксиом используются свойства транзитивности и наследования:

$$r(c_i, c_j) \wedge r(c_j, c_k) \rightarrow r(c_i, c_k) \qquad (A_{tr})$$

$$r(c_i, c_j) \wedge r_1(c_j, c_k) \rightarrow r_1(c_j, c_k) \qquad (A_i)$$

Набор отношений ЛО, их свойства и особенности их описания будут рассмотрены в следующем разделе.

2.2. Модель отношений в ЛО

Для логического вывода при обработке текстов в широкой предметной области необходимо, прежде всего, описывать наиболее существенные отношения между понятиями, сохраняющие свою значимость, надежность в различных контекстах упоминания понятий.

Было выдвинуто предположение, которое подтвердилось в ходе экспериментов в различных предметных областях, что наиболее значимыми отношениями между понятиями являются те, которые связаны с сосуществованием этих понятий или их экземпляров.

В результате в модели лингвистической онтологии используются четыре основных отношения, каждое из которых обладают вышеуказанными свойствами, которые будут подробно рассмотрены в следующих подразделах.

В качестве основных отношений онтологии ЛО используется следующий набор надежных отношений: *выше-ниже*, *часть-целое*, отношение онтологической зависимости, обозначаемое как несимметричная

ассоциация: *асц1-асц2*. Кроме того, в ограниченных случаях используется симметричная ассоциация – *асц*.

2.2.1. Таксономическое отношение *выше-ниже*

Отношение между классами и подклассами понятий может носить разное название в зависимости от терминологических традиций в области использования ресурса: таксономическое отношение, родовидовое отношение, IS-a отношение, отношение гипонимии и гиперонимии (в лексических ресурсах). Данное отношение обладает такими важными свойствами, как транзитивность и наследование, на которых основывается логический вывод во многих компьютерных системах [304]. В данное работе мы будем называть это отношение родовидовым отношением.

Пусть *выше* (c_i, c_j) – родовидовое между понятиями c_i и c_j, c_i является видом (подклассом) c_j, r (c_i, c_j) – это произвольное отношение между понятиями c_i и c_j. Тогда свойства отношения родовидового отношения могут быть записаны следующим образом:

1) транзитивность родовидового отношения

- $(\textit{выше}(c_i, c_j) \wedge \textit{выше}(c_j, c_k) \rightarrow \textit{выше}(c_i, c_k))$

- 2) свойство наследования по родовидовому отношению транзитивность родовидового отношения,

$$\textit{выше}(c_i, c_j) \wedge r(c_j, c_k) \rightarrow r(c_i, c_k)$$

– свойство наследования по родовидовому отношению.

Исторически наиболее ранними принципами установления родовидовых отношений, используемых и в работах по искусственному интеллекту, и в компьютерной лингвистике, было использование ставших классическими диагностических высказываний [38]. Например, если понятие c_i является видом понятия c_j, t_i является текстовым входом понятия c_i, t_j

является текстовым входом понятия c_j можно сказать, что «t_i – *это* t_j», «t_i,.. *и другие* t_j», «*к числу* t_j *относятся* t_i».

Однако позже выяснилось, что одни и те же выражения естественного языка (и, в частности, применяемые диагностические тесты) могут с онтологической точки зрения соответствовать значительно различающимся отношениям между сущностями внешнего мира, в том числе обладающими совсем другими свойствами [68]. Поэтому многие методические руководства по разработке понятийных ресурсов рекомендуют осуществлять дополнительные проверки для устанавливаемого родовидового отношения.

Наиболее распространенной рекомендацией для проверки правильности установления родовидовых отношений является проверка принадлежности экземпляров нижестоящего понятия c_i множеству экземпляров вышестоящего понятия ответ на вопрос: если объект является экземпляром одного понятия, то будет ли он обязательно (т.е. по определению) экземпляром некоторого другого понятия c_j [60, 153, 222], т.е.:

$$\forall \mathrm{c_i}, \mathrm{c_j} \in C, \forall e \in E(\mathrm{c_i}) : \textit{выше}\,(\mathrm{c_i}, \mathrm{c_j}) \rightarrow e \in E(\mathrm{c_j})$$

Критерии проверки правильности установления родовидовых отношений связаны с проверкой выполнения свойств транзитивности и наследования. На проверке транзитивности родовидого отношения основано следующее правило:

> *Нижестоящее понятие и вышестоящее понятие должны относиться к одному и тому же наиболее общему семантическому классу, такому как =действие=, =свойство=, =объект= и т.п.*

Так, стандарты и методические руководства по разработке информационно-поисковых тезаурусов рекомендуют использовать такой принцип для описания иерархических отношений в информационно-поисковых тезаурусах [219].

Второй тип критериев проверки правильности установления родовидовых отношений связан с проверкой свойства наследования. Проверка может носить частный характер, быть связанной именно с конкретной парой понятий. Например, в словарях *изюм* определяется как «*сушеные ягоды винограда*». Следует ли из этого определения, что нужно установить родовидовое отношение *подкласс-класс* между понятиями *ИЗЮМ* и *ЯГОДА ВИНОГРАДА?* С точки зрения наследования свойств ответ на этот вопрос должен быть отрицательным, поскольку изюм не несет многих свойств ягод как плодов некоторого растения: он не растет, не зреет, его не собирают.

Проверка свойств наследования может производиться и на основе общезначимых формальных свойств понятий. Так, для анализа правильности отношений *класс-подкласс* Н. Гуарино и К. Велти [69] предлагают проверять наследование на видовые понятия такого свойства вышестоящего понятия, как «критерий идентичности».

Суть критерия идентичности некоторого понятия заключается в том, чтобы определить, что означает, что две сущности, представляющие примеры одного и того же понятия, являются одним и тем же, как может сущность меняться, сохраняя свою идентичность, какие свойства существенны для сохранения своей идентичности и др. Можно говорить о достаточных условиях идентичности, то есть какие условия используются, чтобы определить идентичность, и о необходимых условиях идентичности, то есть, что следует из того, что два объекта идентичны.

Например, два человека должны быть признаны одним и тем же лицом, если они находились в одном и том же месте в одно и то же время. Таким образом, условием идентичности физических лиц является физическое совпадение нахождения по месту и времени. Если предполагаемое родовое и видовое понятие имеют разные условия идентичности, то это означает, что между ними не может быть установлено отношение *класс-подкласс*.

Одной из серьезных проблем описания родовидовых в онтологиях является их смешение с описанием отношений «тип-роль»: от понятия-типа к понятию-роли.

Понятия-роли занимают «промежуточную» позицию между понятиями-объектами и понятиями-отношениями: роли – это то, что есть, но только в контексте того, что случается. В течение многих лет понятие роли активно обсуждается в таких областях, как концептуальное моделирование и представление знаний. Наиболее часто роль рассматривается посредством двух дополнительных понятий: игрок и контекст. Например, для роли *студент* игроком является человек, а контекст определяется отношением к высшему учебному заведению.

Дж. Сова [190] определяет понятие-роль следующим образом: «Подтипы сущности могут быть двух видов: натуральные типы и ролевые типы, которые являются подтипами натуральных типов в конкретных образцах отношений (particular pattern of relationships). Человек, например, является натуральным типом, а учитель – это подтип человека в ситуации обучения». Сова предлагает простой тест для определения, является ли понятие ролью:

> *rt – является ролевым типом, если сущность может быть охарактеризована как rt только при рассмотрении другой сущности, действия или состояния.*

В соответствии с взглядом Дж. Совы роли ассоциируются с отношениями, но при этом они сущности, а не отношения.

В работе [68] Н. Гуарино отмечает, что тест Совы для различения типов и ролей недостаточен: например, нечто может быть охарактеризовано как автомобиль, только если оно имеет, по крайней мере, колеса и мотор, но автомобиль является типом, а не ролью. В работе [69] условие, сформулированное Совой, заменяется на условие так называемой внешней онтологической зависимости (подробнее см. п. 2.2.2):

Понятие c_i называется внешне зависимым от понятия c_j, если для всех экземпляров c_i должен существовать экземпляр c_j, который не является частью или материалом экземпляра c_i.

Авторы работы [133] замечают, что точнее это условие можно сформулировать так:

Экземпляр понятия c_j не должен быть внутренним для экземпляра понятия c_i, т.е. не должен быть частью, или материалом, или качеством (цвет).

Например, понятие *сын* является внешне зависимым от понятия *родитель*, поскольку существует только в рамках семьи по отношению к своим родителям. С другой стороны, автомобиль не является внешне зависимым от какой-либо сущности, поскольку требует существования мотора, который является частью автомобиля. Таким образом, данное условие формализует определение ролей, данной Дж. Совой.

Вместе с тем, на основе такого определения в класс ролей попадают еще дополнительные сущности такие, как качества: =цвет=, =вес=, =скорость=. Например, если синий – это цвет, то обязательно существует хотя бы один объект, цвет которого синий, при чем этот объект не является частью цвета. Понятие цвета поэтому является зависимым, но не кажется подходящим к понятию роли. Поэтому в работе [69] вводится еще одно условие, которое вместе с условием внешней зависимости дает лучшее определение понятию «роль»:

Понятие c_i является семантически жестким (rigid), если любой экземпляр понятия c_i остается экземпляром c_i в течение всего своего существования.

Например, щенок перестает быть щенком, все еще оставаясь собакой, поэтому собака и животное – это жесткие сущности, а щенок не является жестким понятием.

Таким образом, понятие c_i называется ролью, если оно является внешне зависимым и не является семантически жестким [69].

В соответствии с этим определением качества не могут быть ролями, поскольку они являются семантически жесткими: если цвет прекратит быть цветом, то он станет чем-то еще, потеряет свою идентичность.

Таким образом, оба условия вносят вклад в определение роли. Первое условие описывает, что сущность-роль должна рассматриваться в рамках чего-либо объемлющего, что в определении Дж. Совы называлось *particular pattern of relationships*, второе условие помогает формализовать условие конкретности, особенности, упоминаемое в определении Дж. Совы [190, 191].

Введенные понятия онтологической зависимости и семантической жесткости помогают формализовать понятие натурального типа.

Понятие c_i называется натуральным типом, если оно существенно независимо и семантически жестко.

Таким образом, собака – это пример натурального типа. Цвет не является ни ролью, ни натуральным типом: цвет - семантически жесткий, но является онтологически зависимым. Щенок или хромая собака также не являются ни натуральным типом, ни ролью, поскольку они являются независимыми от внешних сущностей и являются семантически жесткими.

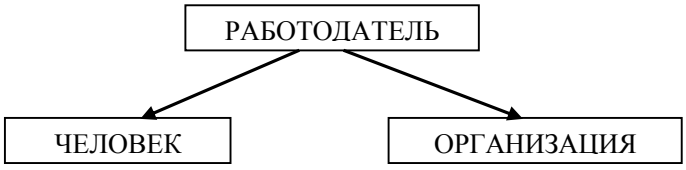


Рис.2.1. Расположение роли над типами сущностей нарушает основной принцип установления родовидовых отношений

Ошибкой при описании ролей является описание их как как вышестоящих понятий для типов, которые могут их занимать. Например,

поскольку работодателем может быть человек или организация, то понятие *РАБОТОДАТЕЛЬ* (рис. 2.1) представляется как вышестоящее, родовое понятие, а понятия *ЧЕЛОВЕК* и *ОРГАНИЗАЦИЯ* представляются как нижестоящие, видовые понятия. Против такого представления выступают многие онтологи [68, 191, 196]. Действительно, такое представление неточно описывает свойства сущностей, поскольку не каждый человек является работодателем, таким образом, нарушается основной принцип установления родовидовых отношений.

В предлагаемой модели лингвистической онтологии для описания ролей используются два основных способа.

Во-первых, если предполагается, что в предметной области большинство примеров того или иного типа будут использованы в некоторой роли, то все-таки устанавливается родовидовое отношение от типа как вида к роли как роду, которое снабжается пометкой V – что означает «возможно по умолчанию». Например, можно установить такое отношение между понятием *СОРБИНОВАЯ КИСЛОТА* и *КОНСЕРВАНТ*, если предполагается, что это основное применение сорбиновой кислоты в данной предметной области, и вероятность встретить в текстах обсуждение сорбиновой кислоты в других применениях (например, в органическом синтезе) в этой области не слишком велика:

выше$_V$ (*СОРБИНОВАЯ КИСЛОТА, КОНСЕРВАНТ)*

Во-вторых, в ЛО может быть введено специальное понятие, обозначающее ролевую принадлежность некоторого типа сущностей. Например, для того чтобы отразить факт, что соли могут выступать в роли электролитов, может быть введено понятие *СОЛЕВОЙ ЭЛЕКТРОЛИТ*. Это целесообразно делать, если такое понятие на самом деле существует и в текстах можно встретить соответствующие языковые выражения:

выше (*СОЛЕВОЙ ЭЛЕКТРОЛИТ, ЭЛЕКТРОЛИТ)*

выше (*СОЛЕВОЙ ЭЛЕКТРОЛИТ, СОЛИ)*

2.2.2. Отношения онтологической зависимости

В предыдущем разделе при обсуждении проблемы отличения ролей от типов было упомянуто отношение онтологической зависимости. Данное отношение играет важную роль в описываемой модели лингвистической онтологии, поэтому в данном разделе это отношение будет рассмотрено более подробно.

2.2.2.1. Определение и подвиды отношения онтологической зависимости

Отношение *онтологической зависимости* между сущностями *X* и *Y* состоит в установлении факта зависимости существования *X* от существования *Y* [119]. Это отношение известно со времен Аристотеля, который заметил, что невещественные сущности, такие, как качества и количества, зависят от вещественных сущностей.

Для выявления онтологической зависимости нужно ответить на следующий вопрос: может ли сущность (*X*) существовать сама по себе, или подразумевает существование чего-либо еще (*Y*). Так, свойство белизны зависит от вещества, например, от куска бумаги, тогда и только тогда, когда это свойство не может существовать без этого куска бумаги.

> ***Определение.*** *X онтологически зависит от Y тогда и только тогда, когда X существует только, если Y существует.*

$D(X, Y) = def$ (*существует* $(X) \rightarrow$ *существует* (Y)).

При описании отношений онтологической зависимости наиболее распространенными и используемыми аксиомами являются аксиомы рефлексивности и транзитивности [53]:

D (X, X) – *рефлексивность отношения зависимости;*

$D\ (X,\ Y)\ \wedge\ D\ (Y,\ Z)\ \rightarrow\ D\ (X,\ Z)$ – *транзитивность отношения зависимости.*

Существует много форм онтологической зависимости.

Если рассматривать онтологическую зависимость конкретной сущности, то можно выделить специфическую, историческую и родовую зависимость [54, 133].

При *специфической зависимости* (SD) конкретная сущность e_1 зависит от другой конкретной сущности e_2, если необходимо, чтобы e_2 существовал, если e_1 – существует:

$$SD\ (e_1, e_2) =_{def}\ ((\exists\ t\ \ pre\ (e_1, t)) \wedge \forall t\ (pre\ (e_1, t) \rightarrow\ pre\ (e_2, t)))$$

где $pre\,(e_i, t)$ – предикат существования сущности e_i в заданное время t [134, 153].

Например, существование конкретного человека зависит от существования его мозга, кроме того, мозг не может быть заменен на другой мозг, т.е. это специфическая зависимость.

При исторической специфической зависимости (HSD) конкретная сущность e_1 зависит от существования другой конкретной сущности e_2 в предшествующий период времени (так дети исторически зависят от своих родителей):

$$HSD\ (e_1, e_2) =_{def}\ ((\exists\ t\ \ pre\ (e_1, t)) \wedge \forall t\ (pre\ (e_1, t) \rightarrow$$
$$\rightarrow \exists t_1((t_1 < t) \wedge\ pre\ (e_2, t_1))))$$

Отношение специфической зависимости между конкретными сущностями может быть естественно перенесено на специфическую зависимость между понятиями (CSD), т.е. понятие c_1 является специфически зависимым от понятия c_2, если все экземпляры c_1 специфически от c_2, т.е.

$$CSD\ (c_1, c_2) =_{def}\ (\forall e_1 \in E(c_1) \exists e_2 (e_2 \in E(c_2) \wedge SD\ (e_1, e_2))))$$

При *родовой зависимости* (generic – GD) существование конкретной сущности зависит от существования конкретных сущностей, относящихся к некоторому понятию *c*:

$$GD\ (e,c) =_{def} ((\exists t\ pre\ (e,t)) \wedge$$
$$(\forall t\ (pre\ (e,t) \rightarrow \exists e_c\ (e_c \in E\ (c) \wedge pre\ (e_c,t))))$$

При родовой зависимости между понятиями (CGD):

$$CGD\ \ (c_1,c_2) =_{def}\ (\forall e_1 \in E(c_1)\ GD\ \ (e_1,c_2))$$

Так, в настоящее время существование конкретного человека зависит от существования его сердца родовой зависимостью, поскольку сердце может быть пересажено, но существование класса человеческих сердец необходимо.

Наконец, могут быть выделены *внутренняя онтологическая зависимость*, т.е. зависимость от внутренних свойств или частей сущности, и *внешняя онтологическая зависимость*, т. е. онтологическая зависимость от существования некоторой отдельной сущности.

В работе [133] указывается еще важность такого вида отношения онтологической зависимости как онтологическая зависимость по определению [49]:

> *сказать, что сущность X зависит от Y, это означает сказать, что Y необходимо (eliminably) должно быть использовано в любом определении X.*

В результате авторы работы [133] уточняют отношение внешней родовой зависимости следующим образом: понятие c_1 является внешне зависимым от понятия c_2, если выполняются два условия:

- определение понятия c_1 необходимо включает понятие c_2;
- если для любой сущности e_1, которая классифицируется как c_1, найдется сущность e_2, которая классифицируется как c_2, являющаяся

внешней для сущности e_1, т.е. не являющейся частью, материалом или свойством сущности e_1.

2.2.2.2. Отношения онтологической зависимости в ресурсах онтологического типа

Отношение онтологической зависимости достаточно часто используется при построении онтологий верхнего уровня.

Для Дж. Совы [191] в построении онтологии верхнего уровня одним из существенных параметров является зависимость понятий друг от друга (prehension). Зависимость может быть внешняя (extrinsic) и внутренняя (intrinsic). Если исчезновение одной из сущности меняет структуру или существование другой сущности, то это отношение внутреннее.

В онтологии BFO (Basic Formal Ontology) [57, 134], целью которой является структурирование онтологий конкретных научных областей, подразделение второго уровня непосредственно связано с понятием онтологической зависимости.

На первом уровне этой онтологии все сущности делятся на сущности, длящиеся во времени, и сущности, происходящие во времени. Длящиеся сущности (Continuants) делятся на пространственные регионы (Spatial regions: пространство, плоскость, прямая, точка), независимые сущности (Independent Continuants), т.е. те, которые могут существовать отдельно от других сущностей, и зависимые сущности (Dependent Continuants), существование которых всегда связано с существованием других сущностей – к таким сущностям относятся функции, роли и качества.

Соответственно, одним из основных отношений онтологии является отношение между зависимыми и независимыми сущностями – *inhere:* c_1 *inheres* in c_2 *at* t, что означает, например, что некоторое качество (например, краснота) зависит от некоторого объекта (например, яблоко или красный жакет), все то время, когда это качество существует, и оно зависит все время своего существования от одного и того же объекта.

В онтологии Dolce [134] отношение зависимости также входит в состав основных отношений онтологии. Разработчики Dolce рассматривают разнообразный набор отношений зависимости:

- специфическая зависимость – зависимость от существования конкретного примера сущности;
- родовая зависимость – зависимость от существования класса сущностей;
- отношения односторонней и двусторонней зависимости;
- отношения пространственной зависимости.

В настоящее время это отношение стало активно обсуждаться и при построении онтологий в области биологии. Так, в работе [92] приводятся примеры отношений онтологической зависимости в области биологии: не может быть клеточного движения без клеток, биологические процессы зависят от органов, клеток и молекул – такая зависимость является специфической. Рассматривая индексирование генных продуктов терминами онтологии GO, авторы работы [25] отмечают, что если продукт проиндексирован термином T_i, и имеется термин T_{i0}, от которого зависит термин T_i, то продукт должен быть проиндексирован и термином T_{i0}.

Особенностью представленной в данной диссертации модели является то, что отношение онтологической зависимости используется для создания лингвистических онтологий в широких предметных областях.

2.2.2.3. Использование отношений онтологической зависимости в информационном поиске

В традиционных информационно-поисковых тезаурусах одним из самых распространенных видов отношений являлось отношение ассоциации, которое было наиболее трудно определить. Некоторые источники излагают наиболее подробные принципы установления ассоциативных отношений, перечисляя разные семантические типы отношений, поскольку в противном случае отношения будут устанавливаться непоследовательно [9].

Американский стандарт Z39.19 [223] описывает наиболее общее правило установления ассоциативного отношения между дескрипторами таким образом, что это отношение сто́ит устанавливать между двумя дескрипторами, если при употреблении одного термина другой термин как бы подразумевается. Более того, один термин часто есть необходимый элемент определения другого термина, например, термин *клетка* составляет необходимую часть определения термина *цитология*. Отметим, что такое правило весьма коррелирует с таким видом онтологической зависимости как зависимость по определению (см. п.2.2.2.1).

Рассмотрим эксперимент, демонстрирующий, с одной стороны, полезность учета отношений онтологической зависимости при информационном поиске, а, с другой стороны, различное поведение разных типов отношений онтологической зависимости при расширении запроса [289]. В качестве таких запросов будем использовать «элементарные» запросы, т.е. запросы, ссылающиеся на одно понятие онтологии. Смысл такого рода элементарных запросов таков: «найти все о c_i», и мы будем обозначать его как SQ (c_i). На практике это означает, что запрос в информационно-поисковую систему, задается посредством однозначного текстового входа, сопоставленного понятию c_i.

Все другие запросы, ссылающиеся на два или более понятий, должны обрабатываться как функция от элементарного запроса. Предполагается, что потенциальное качество расширения запроса на базе отношений онтологии может изучаться на простых запросах. Если поисковые характеристики расширения элементарных запросов являются низкими, то качество расширения сложных поисковых запросов не может быть лучше. Если онтологические отношения дают возможность эффективного расширения запроса для простых случаев, то это является важным шагом для изучения способов расширения сложных запросов.

Рассмотрим два понятия c_1 и c_2, между которыми установлено отношение r. Выполняя простой запрос SQ (c_1), мы хотим узнать, может ли

отношение r с понятием c_2 быть использовано для расширения этого простого запроса. При этом в выдачу по запросу SQ (c_1) с некоторыми весами добавятся документы, содержащие c_2. Следовательно, чтобы проверить полезность такого расширения для запроса SQ (c_1), не нужно выполнять реальное вычисление запроса с расширением, а нужно рассмотреть документы, содержащие c_2, и выяснить, какой процент документов релевантен SQ (c_1).

Мы будем изучать потенциальную эффективность расширения простого запроса для главного понятия c_M в отношении онтологической зависимости текстами, в которых упомянуто зависимое понятие c_D. Для этого были проанализированы 50 первых текстов, полученных по простому запросу SQ(D).

Зависимое понятие c_D	Тип зависимости	Главное понятие c_M	*nD50*	*nM50*
ЛЕС	Специфическая	*ДЕРЕВО*	49	12
САММИТ	Специфическая	*ГЛАВА ГОСУДАРСТВА*	49	20
ПИАНИСТ	Родовая	*ПИАНИНО*	44	16
ГАРАЖ	Родовая	*АВТОМОБИЛЬ*	43	1
АВТОМО-БИЛЬ	Историческая	*АВТОМОБИЛЬНЫЙ ЗАВОД*	18	44

Табл. 2.2. Зависимость качества расширения запроса от типа онтологической зависимости между сущностями.

В качестве запроса задавались слова или выражения, выражающие понятия *ЛО*. Тексты в выдаче упорядочивались на основе стандартной векторной модели *tf.idf* [26]. Поиск был выполнен на коллекции Университетской Информационной Системы РОССИЯ (www.cir.ru), содержащей в момент эксперимента более 800 тысяч документов. Результаты поиска представлены в Табл. 2.2.

Здесь:

- *nD50* – число текстов, содержащих c_D, релевантных c_D и релевантных SQ (c_M),
- *nM50* – число текстов, содержащих c_M, релевантных c_M и релевантных SQ (c_D).

Таблица демонстрирует корреляцию между типом зависимости и поисковыми характеристиками для простых запросов:

- в случае специфической зависимости для практически всех текстов выполняется, что если текст релевантен зависимому понятию, то он релевантен и простому запросу для главного понятия;
- в случае родовой зависимости число текстов, содержащих зависимое понятие и релевантных простому запросу для главного понятия меньше;
- в случае исторической зависимости число текстов релевантных обоим понятиям значительно убывает.

Поисковые характеристики для обратной ситуации в первых четырех случаях (т.е., когда выполняем поиск по главному понятию и смотрим, какие из текстов релевантны зависимому понятию) низки, так как имеется множество текстов, упоминающих главное понятие и не имеющих никакого отношения к зависимому понятию. Одновременно наблюдается отсутствие зависимости понятия c_M от понятия c_D.

В пятой строчке таблицы видно, что значительная доля текстов об автомобильных заводах релевантны простому запросу об автомобилях. При этом нужно заметить, что здесь имеется отношение родовой зависимости: автомобильный завод строится, чтобы выпускать автомобили – имеется отношение родовой зависимости по классу понятия *АВТОМОБИЛЬНЫЙ ЗАВОД* от понятия *АВТОМОБИЛЬ*.

Таким образом, десять вариантов расширения запроса на основе пяти пар понятий показывают важность учета отношений онтологической зависимости при установлении отношение в ресурсе, предназначенном для

информационного поиска, а также демонстрируют различное поведение азных типов отношения онтологической зависимости при расширении простого запроса.

В следующих разделах будет рассмотрено применение теории онтологической зависимости при описании отношений часть-целое и несимметричной ассоциации в предлагаемой модели ЛО.

2.2.3. Отношение *часть-целое*

2.2.3.1. Аксиомы и подвиды отношения *часть-целое*

Вторым типом используемых в ЛО отношений является отношение *часть-целое*, которое играет существенную роль во многих предметных областях.

Особенностью этого отношения является разнообразие его проявлений. При том что наиболее известный подтип отношения *часть-целое* относится к взаимоотношениям между физическими объектами, это отношение может устанавливаться и между сущностями, длящимися во времени, между группами сущностей, ролями и процессами и многое другое [38, 56, 65, 107, 154, 220, 302, 310].

В классической мереологии обычно постулируются три аксиомы для отношения *часть-целое P* [184, 204]:

1) Рефлексивность. Все является частью самого себя:

$$\forall x\, P(x, x)$$

2) Антисимметричность: ничто не является частью своих частей:

$$\forall x\ \forall y\ \ P\ (x,\ y) \wedge P\ (y, x) \rightarrow x = y$$

3) Транзитивность: части частей являются частями целого:

$$\forall x \forall y \forall z\ \ P(x, y) \wedge P(y, z) \rightarrow P(x, z)$$

Данная система аксиом отношения *часть-целое* обычно называется *базовая мереология (ground mereology).*

Поскольку отношение обладает свойством рефлексивности, то выделяются еще строгое отношение *часть-целое*, т. е. когда рассматриваются только части, не равные своему целому PP:

$$PP(x, y) =_{def} P(x, y) \wedge \neg P(y,x)$$

Строгое отношение *часть-целое* является отношением строгого порядка, т.е. выполняются соотношения антирефлексивности, асимметричности и транзитивности:

4) Антирефлексивность

$$\forall x \neg PP(x, x)$$

5) Асимметричность

$$\forall x \forall y \ PP(x, y) \rightarrow \neg PP(y,x)$$

6) Транзитивность

$$\forall x \forall y \forall z \ PP(x, y) \wedge PP(y,z) \rightarrow PP(x,z)$$

Для определенности в дальнейшем в качестве отношения часть целое будет рассматриваться именно строгое отношение *часть-целое PP*.

Другим важным отношением в теории частей [69a, 184, 204] является отношение перекрытия (*overlapping* – •). Отношение перекрытия также включает случай полного вложения одной сущности в другую, а также случай идентичности двух сущностей. Таким образом, это отношение рефлексивно, симметрично, но не транзитивно:

7) Определение отношения перекрытия

$$x \bullet y =_{def} \exists z \ P(z,x) \wedge P(z,y)$$

8) Рефлексивность отношения перекрытия

$$\forall x (x \bullet x)$$

9) Симметричность отношения перекрытия

$$\forall x, y (x \bullet y) \rightarrow (y \bullet x)$$

На основе отношения перекрытия можно сформулировать еще одну аксиому мереологии, называемую принципом слабой дополнительности (weak supplementation principle) [184, 204]:

10) $\forall x, y\, PP(x, y) \rightarrow \exists z\, PP(z, y) \wedge \neg(x \bullet z)$

Теория, включающая аксиомы 1-3 и 10, называется минимальной мереологией. Некоторые авторы ([184]) полагают, что такой набор аксиом представляет собой минимальный набор, которым должна удовлетворять теория частей и целых.

В лингвистике для определения отношения *часть-целое* широко используются лингвистические тесты, т. е. некоторые заданные предложения, в которые подставляются анализируемые сущности. При этом *часть* обычно называется меронимом, а *целое* – холонимом. Естественным тестом для определения меронимии является предложение *X – это часть Y,* которое должно звучать нормально для X и Y, интерпретируемых как родовые понятия*: палец – это часть руки, страница – это часть книги* [38].

Однако многие авторы отмечают, что если применять лингвистические тесты, то возникают серьезные проблемы с транзитивностью отношения *часть-целое,* например, рассмотрим следующую совокупность утверждений: *рука – это часть дирижера*, *дирижер – это часть оркестра,* но странно, если сказать, что *рука – это часть оркестра.*

Рассматривая разные виды отношения *часть-целое*, авторы обычно подчеркивают, что проблемы с транзитивностью связаны со смешением разных видов отношений *часть-целое*. В работе [220] проблемы с транзитивностью объясняются следующим образом: пока используется один тип отношения, то часть-целое всегда транзитивно. Однако когда смешиваются различные отношения меронимии, то возникает проблема с транзитивностью.

В работе [38] подчеркивается, что правильно сформированная иерархия состоит из элементов одного и того же типа. Так, если элемент меронимии – физический объект, то и все другие элементы меронимии должны быть такими же (например, вес тела не должен фигурировать среди его частей). Если один элемент является географической областью, то и другие должны быть такими же (так, Вестминстерское аббатство не является частью Лондона); если один элемент - абстрактное существительное, то и другие должны быть такими же.

В работе [147] предлагается выделить те отношения *часть-целое*, которые, комбинируясь, дают приемлемые результаты транзитивности, и отделить те отношения часть-целое, которые могут привести к ошибочным транзитивным заключениям. Если моделировать такие отношения как *член/коллекция, материал/объект* отношениями, отличными от отношений *часть-целое*, то авторы утверждают, что оставшиеся типы отношений демонстрируют транзитивное поведение, даже если комбинируются произвольным образом. Таким образом, группа отношений компонент/объект, порция/масса, фаза/деятельность, место/местность может быть названа базовыми отношениями *часть-целое*. В рамках любой комбинации базовых отношений *часть-целое* действует правило транзитивности, независимо от комбинации конкретных видов отношений.

Другое мнение высказывается в философской работе [204]. Автор работы утверждает, что проблемы с транзитивностью отношения *часть-целое* и приводимые контрпримеры связаны с неявным сужением понятия «часть» в обыденной речи. То, что ручка двери, являясь функциональной частью двери, может не рассматриваться как функциональная часть дома, не означает, что ручка не является вообще частью дома. Напротив, ручка двери проявляет все обычные свойства частей: масса ручки является частью массы дома; она занимает часть пространства, занятого домом; она будет уничтожена, если уничтожить дом; если уничтожить ручку двери, то и дом будет поврежден.

Если рассмотреть пример: *рука дирижера – дирижер – оркестр,* то также можно видеть, что масса руки является частью массы оркестра, рука дирижера занимает часть пространства, занимаемого оркестром; если будет повреждена рука дирижера, это может вызвать и (может быть, даже серьезные) проблемы с функционированием оркестра.

Сужение понятия «часть» заключается в том, что на интерпретацию понятия «часть» накладываются дополнительные условия (т.е. дополнительное требование, что часть должна быть функциональной и т.п.) и при этом, действительно, свойство транзитивности может не выполняться. В более общем виде, если x – ф-часть (то есть часть с дополнительным условием ф) от y и y – ф-часть от z, x не обязательно является ф-частью от z. Модификатор отношения ф – может не быть транзитивным, но эта ситуация говорит лишь об отсутствии транзитивности у отношения ф-часть, а не у обобщенного отношения в целом.

Помимо классификации отношений *часть-целое* по семантическим основаниям, существуют еще классификации этого отношения по онтологическим свойствам, т.е. на основе анализа сосуществования части и целого [65, 69a].

Так, выделяется отношение *существенной части – EP*. Экземпляр e_1 – является существенной частью экземпляра e_2, если e_2 специфически зависит от e_1:

$$EP\ (e_1, e_2) =_{def}\ SD\ (e_2, e_1) \wedge PP(e_1, e_2) =$$
$$= \forall t(\ pre\ (e_2, t) \rightarrow\ PP\ (e_1, e_2))$$

где PP – предикат *быть частью:* e_1 является частью e_2.

Таким образом, для каждого конкретного человека его мозг является существенной частью. Отметим, что здесь для краткости записи используется предположение, что отношение *часть-целое* обсуждается только для существующих объектов [184], т.е.

$$\forall e_1, e_2, t\ (P\ (e_1, e_2) \rightarrow\ pre\ (e_1, t) \wedge\ pre\ (e_2, t))$$

В современном мире конкретному человеку может быть пересажено другое сердце, но человек обязательно должен иметь сердце. Такое отношение отражается посредством понятия *обязательной части – MP*. Конкретная сущность e_1 является обязательной частью конкретной сущности e_2, если e_1 является экземпляром понятия c, и e_2 имеет родовую зависимость от понятия c:

$$MP\ (c, e_2) =_{def}\ (\forall t\ (pre\ (e_2, t) \rightarrow \exists\ e_1\ (e_1 \in E(c) \wedge PP\ (e_1, e_2)))$$

Часть e_1 называется *неотделимой частью* e_2, если e_1 специфически зависит от e_2, и e_1 является частью e_2:

$$IP(e_1, e_2) =_{def} (\forall t(pre(e_1, t) \rightarrow PP(e_1, e_2)))$$

Примером неотделимой части является мозг человека, который не может существовать вне своего целого. Как уже указывалось, сердце человека может быть отделено от конкретного человека и пересажено другому человеку. Но при этом должна существовать сама категория людей. Таким образом, сердце человека зависит от человека родовой зависимостью, и такая зависимость называется *обязательным целым MW*.

$$MW\ (e_1, c) =_{def}\ (\forall t(pre(e_1, t) \rightarrow \exists\ e_2(e_2 \in E(c) \wedge PP\ (e_1, e_2))))$$

Все эти отношения могут быть перенесены на отношения между понятиями:

Отношение существенной части между понятиями *(CEP)*:

$$CEP\ (c_1, c_2) =_{def}\ (\forall e_2(e_2 \in E(c_2) \rightarrow \exists e_1(e_1 \in E(c_1) \wedge EP(e_1, e_2))))$$

Отношение обязательной части между понятиями *(CMP)*:

$$CMP\ \ (c_1, c_2) =_{def}\ (\forall e_2(e_2 \in E(c_2) \rightarrow \exists e_1(e_1 \in E(c_1) \wedge MP\ (e_1, e_2))))$$

Отношение неотделимой части между понятиями (*CIP*):

$$CIP\ (c_1, c_2) =_{def} (\forall e_1 (e_1 \in E(c_1) \rightarrow \exists e_2 (e_2 \in E(c_2) \wedge IP(e_1, e_2))))$$

Отношение обязательного целого между понятиями (*CMW*)

$$CMW\ (c_1, c_2) =_{def} (\forall e_1 (e_1 \in E(c_1) \rightarrow \exists e_2 (e_2 \in E(c_2) \wedge MW\ (e_1, e_2))))$$

2.2.3.2. Подходы к описанию отношения *часть-целое* в формальных и лингвистических онтологиях

В различных типах онтологических ресурсах были приняты разные решения относительно описания отношения *часть-целое*.

В информационно-поисковых тезаурусах отношения *часть-целое* могут входить в состав иерархических отношений. Иерархические отношения обычно рассматриваются как несимметричные и транзитивные. При установлении иерархических отношений важна независимость от контекста. В частности, в тех случаях, когда имеется множественная принадлежность части к целому, то между такими терминами не должно устанавливаться иерархическое отношение. Между такими дескрипторами может быть установлено отношение ассоциации. Например, карбюраторы являются частями не только автомобилей. Поэтому дескрипторы *КАРБЮРАТОР* и *АВТОМОБИЛЬ* не должны быть связаны отношением *часть-целое* в информационно-поисковом тезаурусе [219].

Если сформулировать это условие с использованием онтологических подвидов отношения *часть-целое,* то можно сказать, что между дескрипторами информационно–поискового тезауруса рекомендуется устанавливать отношение часть-целое, если между соответствующими понятиями существует либо отношение неотделимой части, либо обязательного целого, т.е.:

$$целое_{thes}(c_1, c_2) =_{recom} CIP(c_1, c_2) \vee CMW(c_1, c_2)$$

Впрочем, нужно отметить, что последовательное применение данной рекомендации в каком-либо информационно-поисковом тезаурусе практически не встречается. Для простоты описания отношений *часть-целое* рекомедуется в основном описывать жесткие иерархические системы, как иерархию географических регионов или вложенность военных подразделений [223].

Таким образом, с точки зрения разработки информационно-поисковых тезаурусов не рекомендуется описывать как отношения *часть-целое* такие отношения, как:

- *Сталь – велосипед*, поскольку сталь может быть в разных артефактах, не только в велосипеде;
- *Рука – музыкант*, поскольку руки не только у музыкантов;
- *Кусок – пирог*, поскольку многие другие вещи можно разделить на куски;
- *Дерево – лес,* поскольку деревья растут не только в лесу.

Отметим также, что никаких требований на зависимость целого от части (таких как существенная часть или обязательная часть) в рекомендациях информационно-поисковых тезаурусов не накладывается.

Подход к отношениям *часть-целое* в тезаурусе WordNet принципиально другой, поскольку отношения *часть–целое* устанавливаются в WordNet на основе лингвистического теста:

> *X является частью Y, если можно сказать, что X – это часть Y (An x is a part of Y) или Y имеет X как часть (A y has an x as a part).*

Внутри отношения *часть-целое* дополнительно выделяются отношения *быть_элементом* (человек - часть человечества) и *быть_сделанным_из (стекло – часть стеклянного изделия).* Синсет-часть может быть сопоставлен большому количеству синсетов-целое, как, например, *point* (острие) может быть у *стрелы, ножа, иголки, карандаша, булавки* и т.п.

В различных формальных онтологиях также принимаются различные решения по принципам описания отношений *часть-целое*.

В онтологии SUMO [150] отношения *часть-целое* определены только над осязаемыми (tangible) пространственными сущностями – объектами. Такое ограничение не является типичным для общей мереологии. В этой онтологии отношение *часть-целое* подразделяется на следующие подвиды: член, компонент, кусок (piece), собственно часть, поверхностная часть. Поверхностные части делятся на поверхность, верх, низ и бок.

В онтологии OpenCYC [39] отношение *часть-целое* определяется в очень обобщенном смысле. Единственное ограничение на аргументы отношения заключается в том, что они должны быть конкретными сущностями. Отношение *часть-целое* включает такие подвиды, как пространственные части, временные части, «концептуальные» части (например, содержать_информацию), члены группы и т.п.

В онтологии DOLCE [134] отношение «объект–материал этого объекта» (ваза–глина) рассматривается как отдельное отношение «составляет» (constitute), не являющееся отношением *часть-целое*:

> *X составляет Y тогда и только тогда, когда X может быть субстратом после разрушения Y.*

Такое решение связано с тем, что объект (ваза) и материал, из которого сделан объект, считаются различными сущностями. Если предположить, что между глиной и вазой существует отношение *часть-целое*, то глина должна совпасть с вазой, поскольку у глины и вазы совпадают части, а значит, и по аксиомам мереологии глина и ваза совпадают.

Таким образом, при всей кажущейся очевидности принципов установления отношения *часть-целое,* распространенности этого отношения в различных предметных областях, среди исследователей и авторов ресурсов нет единства в трактовке отношения *часть-целое*.

2.2.3.3. Отношение *часть-целое* в предлагаемой модели ЛО

При описании отношения *часть-целое* в ЛО были сделаны усилия, чтобы обеспечить транзитивность этого отношения.

Если обсуждать свойства транзитивности и наследования для отношения *часть-целое* в ресурсе, предназначенном для автоматической обработки текстов в информационно-поисковых приложениях, то наиболее важной операцией, которую необходимо обеспечить, является релевантность обсуждения частей обсуждению целого. То есть необходимо описывать отношения *часть-целое* так, что если текст или его некоторый фрагмент посвящен обсуждению части, то можно предполагать, что этот текст (или его фрагмент) будет релевантен и обсуждению целого [294].

Важным условием для обеспечения такого наследования является зависимость существования части от существования целого (ср. [11]). Действительно, если все существование некоторой части связано с существованием целого, то и тексты, обсуждающие эту часть, будут иметь непосредственное отношение и к целому, даже если это целое в тексте явно не упомянуто.

Этим требованием, в частности, обеспечивается выполнение рекомендаций руководств и стандартов по разработке информационно-поисковых тезаурусов [219, 223] в том, что описание иерархических отношений должно быть независимо от контекста их упоминания. Описание таких независимых от контекста, «надежных» отношений в ресурсах, предназначенных для автоматической обработки текстов, имеет большое значение, поскольку в автоматическом режиме часто бывает невозможно использовать контекст для подтверждения существования того или иного отношения.

Зависимость части от целого не влечет эксклюзивность части по отношению к целому, т. е. того, что у части ровно одно непосредственное целое. Так, например, локоть является частью руки человека и одновременно

частью костной системы, при этом локоть является зависимой частью и для руки человека, и для костной системы.

Таким образом, при описании отношения *часть-целое* в онтологии применяются принципы, на основе которых в предыдущем разделе формализована рекомендация стандартов по информационно-поисковым тезаурусам: существование экземпляров понятия-части c_1 зависит от существования экземпляров целого c_2 *специфической или родовой онтологической зависимостью,* т.е. экземпляры понятия-части c_1 представляет собой неотделяемые части для экземпляров понятия-целого c_2 или экземпляры понятия c_2 являются обязательным целым для экземпляров c_1:

$$целое_{ло}(c_1, c_2) =_{def} (CIP(c_1, c_2) \vee CMW(c_1, c_2))$$

В качестве аксиом отношения *часть-целое* принимаются аксиомы минимальной мереологии (см. п. 2.2.3.1): *антирефлексивность, асимметричность, транзитивность и принцип слабой дополнительности.*

Возникает вопрос, является ли таким образом определенное отношение $целое_{ло}$ транзитивным. Поэтому необходимо доказать следующее утверждение.

Утверждение 1. Отношение $целое_{ло}$ является транзитивным.

Доказательство. Пусть имеются три понятия c_1, c_2 и c_3, такие что выполняются следующие соотношения: $целое_{ло}$ (c_1, c_2) и $целое_{ло}$ (c_2, c_3); нужно показать, что выполняется также соотношение $целое_{ло}$ (c_1, c_3). Транзитивность отношения *целое* не требует доказательства по принятому набору аксиом. Поэтому нужно доказать, что наложенные ограничения на отношение *часть-целое* являются транзитивными.

Нужно рассмотреть четыре случая.

1) Все экземпляры c_1 специфически зависят от c_2, и все экземпляры c_2 специфически зависят от c_3. Таким образом, любой экземпляр c_1 требует существование конкретного экземпляра понятия c_3, т.е.

является неотделимой частью экземпляра c_3. И, следовательно, в этом случае отношение *целое*$_{ло}$ – транзитивно.

2) Все экземпляры c_1 зависят от c_2 родовой зависимостью, и экземпляры c_2 зависят от c_3 родовой зависимостью. Таким образом, любой экземпляр c_1 требует существование хотя бы одного экземпляра понятия c_2, а существование этого экземпляра понятия c_2 требует существования хотя бы одного экземпляра c_3. Таким образом, для любого экземпляр понятия c_1 экземпляр понятия c_3 является обязательным целым. И, следовательно, и в этом случае отношение *целое*$_{ло}$ – транзитивно.
3) Пусть в этом случае все экземпляры c_1 зависят от c_2 специфической зависимостью, а все экземпляры c_2 зависят от c_3 родовой зависимостью. Таким образом, любой экземпляр c_1 требует существование конкретного экземпляра понятия c_2, а существование этого экземпляра понятия c_2 требует существования хотя бы одного экземпляра c_3. Таким образом, для любого экземпляра понятия c_1 экземпляр понятия c_3 является обязательным целым. И, следовательно, и в этом случае отношение *целое*$_{ло}$ – транзитивно.

И, следовательно, и в этом случае отношение *целое*$_{ло}$ – транзитивно.

4) Наконец, в этом случае все экземпляры c_1 зависят от c_2 родовой зависимостью, а все экземпляры c_2 зависят от c_3 специфической зависимостью. Таким образом, любой экземпляр c_1 требует существование хотя бы одного экземпляра понятия c_2, а существование этого экземпляра понятия c_2 требует существования конкретного экземпляра понятия c_3. Таким образом, для любого экземпляра понятия c_1 требуется существование хотя бы одного экземпляра понятия c_3, т.е. экземпляр понятия c_3 является обязательным целым для экземпляра понятия c_1.

И, следовательно, и в последнем случае отношение *целое*$_{ло}$ – транзитивно.

Утверждение доказано.

Табл.2.1. Возможные классы сущностей в отношении *часть-целое.*

Часть	Целое	Пример	Комментарий
Физический объект	Физический объект	*балкон зала – зрительный зал*	
Место	Место	*Европа – Евразия*	
Вещество	Вещество	*амидная группа - амиды*	
Множество	Множество	*батальон – рота*	
Фрагмент текста	Текст	*строфа – стихотворение*	
Процесс	Процесс	*номер - представление*	
Характерное свойство	Произвольная Сущность	*водоизмещение – судно*	Сходство свойств и частей рассматривается в работах [11, 310]
Роль	Процесс	*инвестор – инвестирование*	Роли как части процессов рассматриваются в [107, 219]
Участник сферы деятельности	Сфера деятельности	*машиностроительный завод – машиностроение*	Похоже на роли

Накладывая вышеперечисленные условия установления отношения *часть-целое*, мы принимаем достаточно широкую трактовку отношения *часть-целое* – см. табл. 2.1.

Таким образом, в настоящее время в ЛО используются следующие свойства отношения *часть-целое:*

часть (c_1, c_2) ↔ *целое* (c_2, c_1)

целое (c_1, c_2) ∧ *целое* (c_2, c_3) → *целое* (c_1, c_3) – транзитивность отношения

выше (c_1, c_2) ∧ *целое* (c_2, c_3) → *целое* (c_1, c_3) – наследование отношения целое по отношению *выше-ниже.*

Приведем примеры вывода на основе свойства транзитивности:

целое (*ОБВИНЯЕМЫЙ ПО ДЕЛУ, СУДЕБНОЕ ОБВИНЕНИЕ*)

∧ *целое* (*СУДЕБНОЕ ОБВИНЕНИЕ, СУДЕБНЫЙ ПРОЦЕСС)*

→ *целое* (*ОБВИНЯЕМЫЙ ПО ДЕЛУ, СУДЕБНЫЙ ПРОЦЕСС)*

целое (АПТЕКА, ЛЕКАРСТВЕННОЕ ОБЕСПЕЧЕНИЕ)

∧ *целое (ЛЕКАРСТВЕННОЕ ОБЕСПЕЧЕНИЕ, МЕДИЦИНСКАЯ ПОМОЩЬ)*

∧ *целое* (*МЕДИЦИНСКАЯ ПОМОЩЬ, ЗДРАВООХРАНЕНИЕ*)

→ *целое (АПТЕКА, ЗДРАВООХРАНЕНИЕ)*

В информационных системах такие цепочки часто интерпретируются следующим образом: если в тексте обсуждается *обвиняемый по делу*, то этот текст релевантен и таким темам, как *судебное обвинение, судебный процесс* и т.д.

В результате в создаваемых по данной модели лингвистических онтологиях реально работает вывод по транзитивности отношений *часть-целое*, что является новым достижением для лингвистических онтологий, поскольку в тезаурусе WordNet транзитивность отношения *часть-целое* не

предполагалась, а в рекомендациях по информационно-поисковым тезаурусам это отношение сводилось к весьма узкому набору случаев, из-за чего такой вывод не мог играть значительной роли.

2.2.4. Отношение внешней онтологической зависимости в моделиЛО

В предыдущем разделе мы предложили использовать для описания внутренних характеристик и частей класса понятий отношения специфической и родовой онтологической зависимости. В данном разделе мы рассмотрим тип отношений, предлагаемых нами для описания отношений классами сущностей, существующих отдельно.

Если рассмотреть примеры специфической и родовой зависимости между внешне существующими понятиями, то можно видеть, что такие отношения не являются полезными для обработки текстов рамках информационного поиска. Так, человек зависит родовой зависимостью от кислорода и многих других сущностей, которые обеспечивают существование живых существ на планете Земля.

В результате, после многих экспериментов был сделан вывод, что в онтологии, предназначенной для автоматической обработки текстов, прежде всего, для приложений информационного поиска, необходимо, прежде всего, отражать внешнюю родовую зависимость (см. п. 2.2.2.1), т.е. зависимость существования экземпляров понятия от существования другого понятия, например, гараж зависит от автомобиля внешней родовой зависимостью. Поскольку гараж как постройка не перестанет существовать, если в мире исчезнут все автомобили, но ее свойство «быть_гаражом» зависит от существования класса сущностей «автомобили».

Отношение внешней родовой зависимости является несимметричным, и для его обозначения используется отношение несимметричной ассоциации $асц_1 – асц_2$. Отношение $асц_1$ ведет от зависимого понятия к главному понятию

отношения родовой зависимости, а отношение acu_2 является к нему обратным отношением.

Таким образом, отношение несимметричной ассоциации является отношением внешней родовой зависимости и устанавливается между понятиями c_1 и c_2 при одновременном выполнении следующих условий:

1) Родовая зависимость понятия c_1 от c_2:

$$CGD\ (c_1, c_2) =_{def} (\forall e_1 (e_1 \in E\ (c_1) \rightarrow \exists t (pre\ (e_1, t) \wedge$$
$$\wedge (\forall e_1, t\ (pre\ (e_1, t) \rightarrow (\exists e_2 \in E\ (c_2) \wedge pre\ (e_2, t)))))$$

2) Внешняя онтологическая зависимость понятия c_1 от c_2 в виде условий 2а) и 2б):

2а) Отношение между понятиями c_1 и c_2 не может быть представлено как *часть-целое*:

$$\neg\ часть\ (c_1, c_2) \wedge \neg\ часть\ (c_2, c_1)$$

2б) Отношение между понятиями c_1 и c_2 не может быть представлено как отношение *часть-целое* вышестоящего понятия:

$$\neg\exists\ C_k : выше\ (c_1, c_k) \wedge (часть\ (c_k, c_2) \vee часть\ (c_2, c_k))$$

3) Применение диагностического теста вида "Существование понятия c_1 требует существования понятия c_2", которое должно восприниматься как истинное в рамках системы понятий заданной предметной области.

Это отношение формализует рекомендацию стандартов и руководств по созданию информационно-поисковых тезаурусов (см. например [223]), которые указывают на важность анализа определений для установления ассоциативных отношений в информационно-поисковых тезаурусах. Новое в предложенном подходе заключается в следующем.

Во-первых, выделена совокупность отношений, которые описываются как отношения *часть-целое.* Во-вторых, более формализованное отношение

внешней родовой зависимости позволяет не полагаться на имеющиеся определения, которые могут быть неполными или избыточными, а проводить анализ самостоятельно.

В настоящее время в приложениях используются следующие свойства отношения внешней родовой зависимости, обозначаемой как несимметричная ассоциация:

$асц_1 (c_1, c_2) \leftrightarrow асц_2 (c_2, c_1)$

Наследование отношения несимметричной ассоциации на виды и части:

$выше (c_1, c_2) \wedge асц_1 (c_2, c_3) \rightarrow асц_1 (c_1, c_3)$

$целое (c_1, c_2) \wedge асц_1 (c_2, c_3) \rightarrow асц_1 (c_1, c_3)$

Условия транзитивности на данное отношение несимметричной ассоциации не накладывается, несмотря на то, что для отношения онтологической зависимости транзитивность обычно постулируется (см. п. 2.2.2). Это связано со сложностью накладываемых ограничений, которые могут оказаться нетранзитивными.

2.2.5. Отношение симметричной ассоциации

Существует несколько ситуаций, когда оправданно представление отношений между понятиями в виде симметричной ассоциации. При этом предполагается, что степень ассоциации между понятиями достаточно высокая, т.е. если два понятия c_1 и c_2 связаны отношением симметричной ассоциации, то тексты, содержащие понятие c_1 часто релевантны запросам, выражающим понятие c_2, и наоборот.

Симметричные ассоциации используются для отражения отношения между понятиями, которые являются взаимозависимыми, но между которыми невозможно поставить отношение *часть-целое*, например,

РОДИТЕЛИ – ДЕТИ (*СЫНОВЬЯ И ДОЧЕРИ*)

Симметричной ассоциацией описывается также отношение между близкими по смыслу понятиями, относящимися к одному и тому же родовому понятию, текстовые входы которых используются как квазисинонимы. Например, есть близкие понятия *АВИАЦИОННАЯ МЕДИЦИНА* и *КОСМИЧЕСКАЯ МЕДИЦИНА*, также имеется множество контекстов употреблений словосочетаний *авиакосмическая медицина, авиационная и космическая медицина.* В некоторый момент развития лингвистической онтологии отношение между такими понятиями может быть отражено в виде симметричной ассоциации.

Наконец, некоторые виды антонимов могут быть представлены в лингвистической онтологии в виде симметричной ассоциации между соответствующими понятиями. Отношением симметричной ассоциации представляются обычно отношения между антонимами, содержащими указание на разную степень, меру одного и того же качества, свойства

В настоящее время используются следующие свойства отношения симметричной ассоциации:

$асц\ (c_1,c_2) \rightarrow асц\ (c_2,c_1)$ – Симметричность отношения

Наследование отношения ассоциации на виды и части:

$выше\ (c_1,c_2) \wedge асц\ (c_2, c_3) \rightarrow асц\ (c_1, c_3)$

$целое\ (c_1,c_2) \wedge асц\ (c_2, c_3) \rightarrow асц\ (c_1, c_3)$

2.3. Группировки понятий и отношений в ЛО.

Для различных приложений автоматической обработки текстов применяются некоторые группировки понятий (окрестности) и отношений (пути) в ЛО.

Для каждого понятия $c_i \in C$ может быть определена окрестность понятия $O_i \subset C$, такая, что $c_j \in O_i$, если существует набор понятий $\{c_1, \ldots c_k\}$ такой, что $r_1\ (c_i, c_1), \ldots r_2\ (c_n, c_{n+1}) \ldots r_r\ (c_k, c_j) \in R$, и на основе аксиом A_t, A_i выводимо отношение $r\ (c_i, c_j)$:

$r_1 (c_i, c_1), \ldots r_2 (c_n, c_{n+1}) \ldots r_r (c_k, c_j) \mapsto r (c_i, c_j)$

На рис. 2.2 изображена схема окрестности понятия.

На множестве отношений ЛО может быть введено отношение иерархии I по следующим правилам:

выше $(c_1, c_2) \rightarrow I (c_1, c_2)$

целое $(c_1, c_2) \rightarrow I (c_1, c_2)$

асц$_1$ $(c_1, c_2) \rightarrow I (c_1, c_2)$

асц $(c_1, c_2) \rightarrow I (c_1, c_2)$

Это отношение означает, что правый элемент отношения считается более высоким по иерархии, чем левый. Для отношения симметричной ассоциации оба члена отношения равноправны.

В окрестности понятия c_i можно определить верхнюю полуокрестность O^+ и нижнюю полуокрестность O^-:

$O^+(c_i) \cup O^-(c_i) = O (c_i)$

$c_j \in O^+ (c_i)$, если $c_j \in O (c_i) \wedge I (c_{i,} c_j)$

$c_j \in O^- (c_i)$, если $c_j \in O (c_i) \wedge I (c_{j,} c_i)$

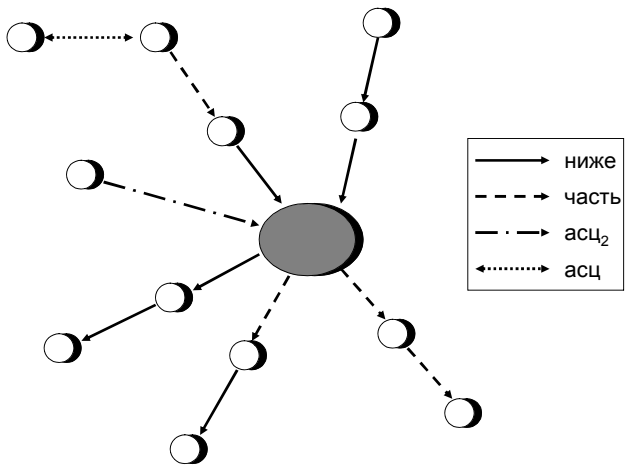


Рис. 2.2. Схема окрестности понятия ЛО.

Пересечение $O^+(c_i)$ и $O^-(c_i)$ может быть непустым из-за существования отношений симметричной ассоциации, входящих в обе полуокрестности. Верхняя полуокрестность понятия c_i также называется *дерево-вверх* понятия c_i, нижняя полукрестность понятия c_i – *деревом-вниз* понятия c_i.

Можно определить следующие виды путей между понятиями:

- *Путь по иерархии вверх* P_{up} (c_0, c_{00}): От понятия c_0 к понятию c_{00} существует путь по иерархии вверх, если $c_{00} \in O^+(c_0)$;

- *Путь по иерархии вниз* P_{down}(c_0, c_{00}): От понятия c_0 к понятию c_{00} существует путь по иерархии вниз, если $c_{00} \in O^-(c_0)$;

- *Путь с перегибом вверх* P_{updown}(c_0, c_{00})*:* Между понятиями c_0 и c_{00} такими, что $c_0 \notin O(c_{00})$ и $c_{00} \notin O(c_0)$, существует путь с перегибом-вверх, если существует точка перегиба – понятие c_i такое, что:

$$\exists\, c_i : c_i \in O^+(c_0) \wedge c_i \in O^+(c_{00})$$

- *Путь с перегибом вниз* P_{downup}(c_0, c_{00}): Между понятиями c_0 и c_{00} такими, что $c_0 \notin O(c_{00})$ и $c_{00} \notin O(c_0)$, существует путь с перегибом-вниз, если существует точка перегиба – понятие c_j такое, что:

$$\exists\, c_j : c_j \in O^-(c_0) \wedge c_j \in O^-(c_{00}).$$

Введенные типы концептуальных путей используются в процедурах автоматического разрешения лексической неоднозначности, расширения поискового запроса, вывода рубрик по тексту.

2.4. Лингвистические онтологии, созданные на основе описанной модели

Вышеописанные принципы были положены в основу разработки нескольких больших ресурсов для информационного поиска: Общественно-политического тезауруса [274, 278], тезауруса русского языка РуТез [110, 280, 286], онтологии по естественным наукам и технологиям ОЕНТ [257, 261], Тезауруса Банка России, Авиа-онтологии [43, 253] и ряда других.

К началу 2012 года объем тезауруса РуТез составляет 53 тысячи понятий, 153 тысячи текстовых входов, более 210 тысяч отношений между понятиями, более 1 млн. отношений, выводимых по установленным правилам вывода. Объем онтологии ОЕНТ также составляет более 50 тысяч понятий, 140 тысяч терминов. Величина Общественно-политического тезауруса – 38 тысяч понятий, более 100 тысяч текстовых входов.

Вышеперечисленные ресурсы имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними, которые устанавливаются в соответствии с требованием правомочности расширения запроса по иерархии связей при информационном поиске. Эти ресурсы принадлежат к особому классу онтологий, так называемым лингвистическим онтологиям, поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса. В то же время они являются тезаурусами, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте, – такой набор текстовых входов понятий необходим для использования онтологий для автоматической обработки текстов.

Разнообразие предметных областей, для которых созданы эти ресурсы, доказывают универсальность предложенной модели лингвистической онтологии, т.е. посредством такой модели можно описывать базовые свойства и отношения понятий, присутствующие в любой предметной области. Объемы созданных ресурсов демонстрируют удобство модели для быстрого наращивания ресурсов.

Особенности построения тезаурусов для автоматического концептуального индексирования и технологии использования их в различных приложениях информационного поиска рассмотрим на примере Общественно-политического тезауруса.

Общественно-политическая область включает в себя лексику и терминологию, которая, с одной стороны, известна достаточно широким слоям населения, с другой стороны, соответствует понятиями профессиональных сфер деятельности (см. также концепцию «универсального терминологического пространства в [300].

На такую особенность Общественно-политической области указывают также разработчики Тезауруса Исследовательской службы Конгресса США [106], которые пишут, что для описания широкой области общественных отношений приходится использовать разные типы лексических единиц, в том числе, как специальную терминологию, так и тематическую лексику общего языка (popular terminology).

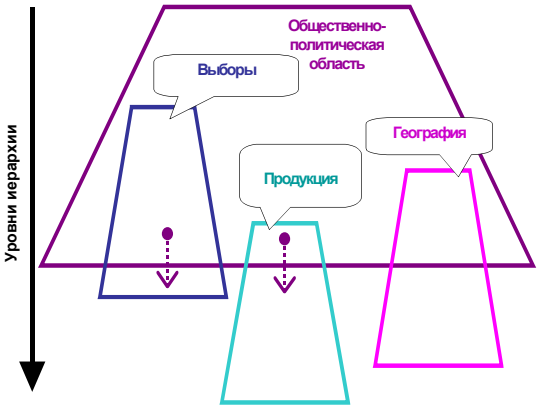


Рис.2.1. Специальная лексика vs. Общественно-политическая

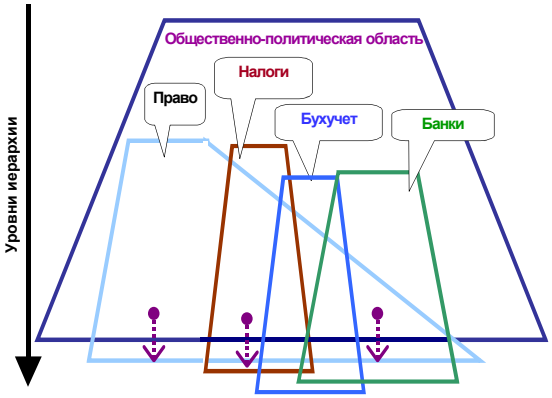


Рис.2.2. Взаимосвязь специальной лексики внутри Общественно-политической области

Разные предметные области имеют различные по величине перенесению с общественно-политической областью. Так, понятийная система предметной области «Выборы» практически полностью находится в Общественно-политической области, в то время как сферы различных промышленных производств пересекаются с общественно-политической областью лишь по небольшому числу понятий (рис. 2.1). Можно выделить совокупность непроизводственных регулирующих сфер деятельности, которые значимы в повседневной деятельности многих людей и, значит, в

значительной степени пересекаются с общественно-политической областью, такие, как Налоги, Бухгалтерия, Право, Таможня, Банковская сфера, образуя правовой и финансовый блоки областей (рис. 2.2).

Выделение такой области, а также выделение среди общеупотребительной лексики слов, принадлежащих этой области, является чрезвычайно полезным для разработки лингвистических ресурсов и технологий автоматической обработки больших электронных коллекций.

Прежде всего, терминология и лексика из этой области активно используется в самых разных по жанру, но значимых для жизни общества текстах, как законы, международные договора, другие официальные документы, газетные сообщения, экономические документы [112]. Таким образом, создание лингвистического ресурса в общественно-политической области может значительно повысить эффективность и содержательность обработки всех этих видов документов.

Поскольку общественно-политическая область содержит наиболее общеизвестные понятия многих профессиональных предметных областей, то лингвистический ресурс, разработанный для общественно-политической области, может стать источником существенного понятийно-терминологического материала для создания лингвистических ресурсов в конкретных предметных областях.

Одновременно общественно-политическая область – это область общезначимая и содержит значительное количество общелексического материала, который относится к нижним и средним наиболее конкретным уровням языковой системы языка, поэтому понятийная структура общественно-политической области является и существенным базисом, на который можно опираться, например, выстраивая понятийную иерархическую систему языка типа WordNet [138].

Кроме того, если рассмотреть количество многозначных общезначимых слов внутри общественно-политической области и в общем лексиконе, то многозначных слов в общественно-политической области

значительно меньше, а процедура автоматического разрешения многозначности работает эффективнее, поскольку часто значения относятся к различным подобластям общественной жизни, например, в подавляющем большинстве текстов контексты разных значений словоформы *судов* как средства водного транспорта и судебного органа существенно различаются. Это различие можно также эффективно использовать при автоматической обработке текстов, используя, например, комбинированную обработку текстов и запросов при решении информационно-поисковых задач, а именно пытаться разрешать многозначность для слов и терминов, относящихся к общественно-политической области, и использовать пословную обработку для остальной общеупотребительной лексики.

В настоящее время Общественно-политический тезаурус интегрирует в себе значительную долю терминологии следующих предметных областей, которая была введена в него в течение деятельности в ряде проектов по автоматической обработке текстов: экономика, право, социология, демография, банковское дело, государственный финансовый контроль, выборы.

Общественно-политический тезаурус может рассматриваться как информационно-поисковый тезаурус, созданный для автоматического индексирования текстов в широкой общественно-политической области. По широте предметной области Общественно-политический тезаурус соответствует таким тезаурусам как Тезаурус исследовательской службы Конгресса США LIV [106] или тезаурус Европейского сообщества EUROVOC [265]. Однако Общественно-политический тезаурус, созданный по описанной выше модели, во много раз больше упомянутых тезаурусов.

Такое различие связано с тем, что Общественно-политический тезаурус изначально создавался как ресурс для автоматической обработки текстов, когда человека-посредника между информационно-поисковым тезаурусом и языком документов нет. Поэтому достаточно большой объем информации должен быть представлен непосредственно в тезаурусе.

Общественно-политический тезаурус включает не только термины, которые представляют важные понятия в текстах данной предметной области, но также охватывает широкий круг более специфических терминов, обнаружение которых в конкретном тексте сделает этот текст релевантным запросу по понятиям более высокого уровня.

Синонимические ряды понятий Общественно-политического тезауруса значительно богаче, чем совокупности вариантов дескриптора в тезаурусах LIV или EUROVOC, поскольку синонимы должны описывать различные способы выражения данного понятия в тексте для автоматического процесса, а не для человека. Ряды синонимов включают в себя не только существительные и именные группы, а также прилагательные, глаголы, глагольные группы.

Расширение терминологической базы Общественно-политического тезауруса ведет к необходимости описания многозначных терминов. Общественно-политический тезаурус содержит около 4.5 тысяч многозначных слов и выражений. В традиционных информационно-поисковых тезаурусах нет необходимости аккуратно описывать многозначность употребляемых в текстах слов и выражений, поскольку понимание текста, его основной темы возложено на человека-индексатора. Расширение понятийной базы Общественно-политического тезауруса ведет к увеличению и усложнению функций отношений между понятиями тезауруса: возникает необходимость логического вывода на отношениях.

Заключение к главе 2

В данной главе мы представили модель лингвистической онтологии для автоматической обработки текстов, учитывающую три существующие методологии создания компьютерных ресурсов. Существенно новым в предложенной модели является набор отношений лингвистической онтологии, который специально подобран для описания широкой предметной области.

Для качественного выполнения всех различных функций отношений ЛО при автоматической обработке текстов в приложениях информационного поиска важно обеспечить многошаговый логический вывод, что может быть достигнуто на базе свойств транзитивности и наследования. Кроме того, при описании отношений необходимо добиться того, чтобы отношения были максимально «надежными», не зависели от контекста упоминания понятия.

Для обеспечения этих свойств было предложено использовать небольшой набор отношений, сопоставимый с набором отношений в традиционных информационно-поисковых тезаурусах. Однако были введены более строгие онтологические определения используемых отношений. Такая система отношений отражает наиболее существенные взаимосвязи между сущностями, может применяться для описания отношений между понятиями в самых разных предметных областях.

Разнообразие предметных областей, для которых созданы лингвистические онтологии по предложенной модели доказывает универсальность этой модели, ее способность описывать базовые свойства и отношения понятий, присутствующие в любой предметной области. Объемы созданных ресурсов демонстрируют удобство модели для быстрого наращивания ресурсов.

Глава 3. Лингвистическая онтология как средство моделирования структуры связного текста

Как уже указывалось, распространенной моделью обработки связного текста в информационных системах является модель мешка слов (bag of words), когда предполагается, что все слова в тексте употребляются независимо друг от друга, и, таким образом, значимость слова в тексте определяется как функция от особенностей употребления в тексте именно этого слова, прежде всего, от частоты его употребления в этом тексте.

Такая модель противоречит известной особенности связного текста, заключающейся в том, что если текст посвящен какой-то теме, то в нем употребляется множество слов и выражений, относящихся к этой теме. Наличие большого лингвистико-онтологического ресурса позволяет выявить взаимоотношения между словами, и, таким образом, задача состоит в том, чтобы найти модель, которая позволяла бы эффективно преобразовывать информацию о взаимосвязи слов в оценку значимости конкретных слов в тексте, и использовать полученные оценки в приложениях автоматической обработки текстов

Таким образом, необходимо понять выяснить, что представляет собой эта совокупность близких по смыслу текстовых единиц, и как наличие той или иной совокупности таких единиц влияет на оценку значимости слова. Для создания такой модели необходимо рассмотреть основные особенности связных текстов.

3.1. Моделирование структуры связного текста

Многие модели обработки текстов в сфере информационного поиска базируются на предположении о независимом употреблении слов (bag of words models) в связном тексте. Между тем известно, что текст содержит

множество связанных по смыслу слов, а также имеет внутреннюю иерархическую структуру.

Существует достаточно много разных приложений автоматической обработки текстов, которые могли выдавать более качественные результаты, если бы можно было бы автоматически выявлять содержательную структуру связного текста. Среди них такие приложения, как автоматическое сегментирование текстов, разрешение многозначности, собственно информационный поиск, более качественное определение весов термов в документе, рубрикация текстов, автоматическое аннотирование текстов и др.

Понятие связности текста может быть рассмотрено в нескольких аспектах.

Выделяют когезию или структурную связность и когерентность текста. Фактически речь идет о внутренней (структурной) и внешней (прагматической) связности. Когезией называется связь элементов текста, при которой интерпретация одних элементов текста зависит от других [37, 269]. Когерентностью называется связность, привносимая чем-то внешним по отношению к тексту, прежде всего знаниями его адресата. На основании этих знаний адресат может конструировать определенный ожидания и достраивать связи, отсутствующие в тексте в явном виде [29, 128, 145, 240, 269, 270, 313].

С другой точки зрения выделяют глобальную и локальную связность текста. Глобальная связность текста обеспечивается тем, что у текста имеется единая тема. Локальная связность дискурса проявляется во взаимосвязи между соседними минимальными единицами текста [42, 248].

3.1.1. Тематическая структура и тематическая связность текста

Определение основной темы текста является важным этапом для многих приложений информационного поиска. Понятие основной (или глобальной) темы текста связано с такими свойствами текста как тематическая связность и тематическая структура. Текст может быть

формально связным посредством различных типов связности, но если у него нет единой темы, то он не может рассматриваться как текст [305].

Автор работы [199] указывает на различие трактовки термин *глобальная тема* у разных авторов. Этот термин может относиться к наиболее центральному участнику ситуации, описываемой в тексте. Также термин *глобальная тема* относится к тому, чему посвящен весь текст – и тогда глобальная тема скорее пропозиция, а не именная группа.

Авторы работы [20] предлагают называть главный персонаж, объект, идею термином «тематический элемент» (topic entity) и отделять понятие тематического элемента от термина *глобальная тема текста*. Именно так мы и будем употреблять термины *главная (или основная) тема документа* и *тематический элемент* или *элемент главной темы* документа.

Гипотеза, лежащая в основе многих работ, заключается в том, что содержание текста может быть представлено в виде иерархической структуры пропозиций [42, 199, 240, 266, 303, 313], самая верхняя пропозиция собственно и представляет собой основную тему документа, а пропозиции нижних уровней представляют собой локальные или побочные темы документа.

Ван Дейк [42] описывает тематическую структуру текста, макроструктуру как иерархическую структуру в том смысле, что тема целого текста может быть описана как единственная макропропозиция. Тема целого текста может быть охарактеризована в терминах подтем, а подтемы в терминах еще более локальных подтем. Каждое предложение текста соответствует той или иной подтеме иерархической структуры текста.

Макроструктура текста определяет его глобальную связность. «Без такой глобальной связности, невозможно было бы осуществлять контроль за локальными связями (local connections and continuations). Предложения могут быть хорошо связанными между собой в соответствии с критериями локальной связности, но они могли бы отклониться в сторону, если бы не было глобальных ограничений на их содержание» [42: стр.115-116].

Даже при ручной обработке текстов экспертами трудно последовательно учитывать иерархическую структуру текста. Так, при ручном индексировании или рубрицировании документов разная трактовка побочных тем документа разными экспертами является одним из существенных факторов субъективности этих процессов.

При автоматической обработке документов важность слова или термина для содержания текста, их близость к основной теме документа оценивается с помощью специальных весов. Предполагается, что чем выше в иерархии тематической структуры упомянуто слово или термин, чем ближе они к основной теме документа, тем больше должна быть величина присвоенного веса. Самой простой характеристикой моделирующей такой вес естественно является величина частоты употребления слова (термина) в документе, а также различные ее модификации.

3.1.2. Когезия как структурная связность текста

Еще одним видом связности в тексте является когезия, представляющая собой совокупность лексических и грамматических средств для выражения связей между единицами текста. Когезия может выражаться в тексте несколькими разными способами [70, 240, 269]. Частым видом когезии, которая может моделироваться посредством лексических и онтологических ресурсов, является лексический повтор или лексическая связность. Авторы известной работы [70] классифицируют лексическую связность на пять категорий:

- повторение – употребляется одно и то же слово;
- синонимическое повторение;
- связность через обобщение или специализацию (родовидовые отношения);

- связность через отношения *часть-целое*, например, *Детский сад откроют не раньше понедельника. Еще предстоит просушить все комнаты* (комнаты как часть детского сада);
- связность через коллокацию, сюда же относится антонимия. Такие отношения могут быть выявлены путем статистики частого совместного упоминания слов. Последние четыре вида лексической связности могут быть названы семантическим повтором.

Многие авторы указывают, что лексическая связность – это не просто связи между парами слов текста, а достаточно длинные цепочки слов, близких по смыслу [76, 145, 203, 269]. Так, М. Кронгауз [269] пишет, что средством когезии является вообще подбор тематической лексики, т.е. лексики, относящейся к одному семантическому полю, и соответственно повтор в тексте интегральных признаков этого поля. В работе [145] указывается, что лексическая связность возникает не только между парами слов, но связывает между собой группы слов текстового фрагмента, посвященного одной и той же теме. В работе [267] рассматриваются цепочки семантически связанных слов в стихотворных текстах такие, как «вечер», «утро», «час», «секунда» (имеют семантический признак «время»); «мир», «даль», «расстояние»; «поезд», «путь», «движение»; «тело», «рука», «глаза»; «открытка», «поздравление», «привет» (семантический признак «расстояние»).

В работах [73, 146] рассматривается понятие «гармонии связности», посредством которого делается попытка формализовать отношения внутри предложения и между предложениями. Гармония связности базируется на цепочках когезии, в том числе лексических цепочках, и семантических отношениях, таких, как *агенс, объект, инструмент*, между элементами разных цепочек, устанавливаемыми внутри предложений. R. Hasan [73] указывает, что два языковых выражения должны рассматриваться как единицы одной цепочки, если они более чем один раз выступали в одном и том же отношении в рамках какой-либо ситуации или по отношению к какой-

либо третьей сущности. Подчеркивается, что единство текста основывается на том, что «похожие вещи говорятся о похожих или тех же самых сущностях или событиях. Тексты, в которых больше сущностей участвуют в гармонии связности, рассматриваются людьми как более связные.

3.2. Моделирование лексической связности на основе тезаурусов

Первой работой, в которой предлагалось использовать имеющиеся тезаурусы для автоматического выявления лексической связности текста в виде лексических цепочек и были предложены алгоритмы построения лексических цепочек на основе тезауруса Роже, была работа [145].

В работе указывалось, что лексическая связность возникает не только между парами слов, но связывает между собой группы слов текстового фрагмента, посвященного одной и той же теме. По определению авторов работы лексическая цепочка – это последовательность слов текста, в которой каждое следующее слова связано некоторым отношением с предшествующими словами цепочки. Лексические цепочки не останавливаются на границах предложений и могут проходить через целый текст. Авторы работы рассматривают лексические цепочки как важный шаг на пути к построению риторической и тематической структуры дискурса.

Эксперименты с использованием тезауруса Роже проводились вручную, поскольку на тот момент не существовало электронных версий тезауруса. С появлением тезауруса WordNet подавляющее число экспериментов по построению лексических цепочек было проведено с помощью этого тезауруса.

Первой опубликованной работой, в которой в качестве ресурса для построения лексических цепочек использовался WordNet, была работа [76]. Авторы предполагали использовать лексические цепочки для обнаружения малапропизмов, т.е. ошибок текста, в которых ошибочно написанное слово оказывается реально существующим словом языка, что и затрудняет обнаружение ошибки [237].

Для построения лексических цепочек все отношения между словами, которые могут быть индикаторами лексической связности, делятся на три группы: экстра-сильные, сильные и средней силы. Экстра-сильные отношения устанавливаются только между буквальными повторами слов.

Сильные отношения устанавливаются в трех случаях:

- когда два слова описаны как синонимы (*human* и *person*);
- когда два слова связаны горизонтальным отношением (антонимия, подобие);
- если многословное выражение – единица WordNet – включает в себя однословное (*school – private school*).

Сильное отношение имеет меньший вес, чем экстра-сильное, и больший вес, чем отношение средней силы.

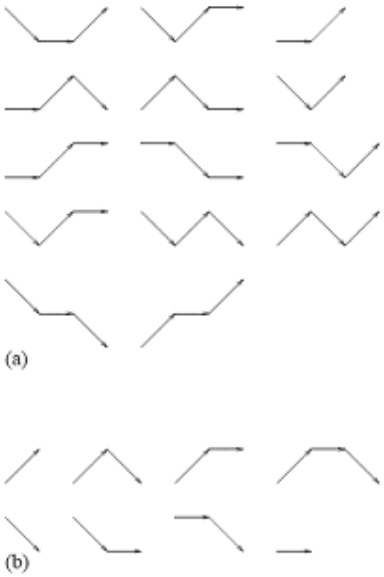


Рис. 3.1. По [76] а) запрещенные пути при построении отношений средней силы, б) разрешенные пути отношений средней силы

Отношения средней силы возникают, когда имеется путь заданной формы между понятиями, к которым относятся два слова. Максимальная длина пути – пять отношений. Не позволяется поворот пути «вниз-вверх».

Разрешен только один поворот «вверх – вниз» и два поворота пути следующего вида: «вверх – горизонтально – вниз». Таким образом, помимо повторов и синонимов рассматриваются как способные участвовать в образовании лексической связности текста следующие слова:

- слова, являющиеся нижестоящими или частями одного и того же понятия от 1 до 4 уровней;
- слова, лежащие на одной иерархической линии гиперонимов, отношений целое, смешанных отношений гипероним-целое в различных вариантах (см. рис. 3.1).

Предполагается, что лексическая связность текста моделируется совокупностью лексических цепочек слов, чьи значения близки по смыслу. Для выявления этих цепочек предлагается следующий алгоритм:

1) текст просматривается пословно с начала до конца. Просматриваются только существительные.
2) первое слово создает первую лексическую цепочку.
3) для каждого следующего слова проверяется, связано ли оно какими-либо лексически-существенными связями с предшествующими словами (и соответственно, лексическими цепочками):
   - если нет, то слово образует новую цепочку;
   - если очередное слово связано только с одной лексической цепочкой, то туда оно и присоединяется;
   - если очередное слово связано с несколькими лексическими цепочками, то выбирается наиболее сильная связь. Выбирается всегда одна лексическая цепочка.
4) в процессе такого построения цепочек происходит разрешение многозначности слов, поскольку значения, по которым не было подсоединения к существующей цепочке, удаляются.

Кроме того, имеются ограничения просмотра – 7 предложений для сильных связей и 3 предложения для связей средней силы.

Авторы данной работы предполагали построить детектор малапропизмов, используя следующую гипотезу: слова, которые не формируют лексические цепочки с другими словами текста, являются потенциальными малапропизмами, поскольку они как бы не соответствуют содержанию текста. Если такое слово обнаруживается, алгоритм подыскивает слова, которые близки по написанию к данному слову и которые удается присоединить к одной из существующих лексических цепочек. Тот вариант, который сильнее всего оказался связанным с существующей лексической цепочкой, считается правильным, т.е. именно тем исходным словом, в котором произошла ошибка.

Авторы протестировали свой подход на материале 500 статей Wall Street Journal, в которые были специально внесены малапропизмы, в среднем один малапропизм на 200 слов – всего 1409. Эксперименты показали точность выявления малапропизмов – 12.5% и полноту 28.7%. В дальнейшем Буданицким [22] было показано, что обнаружение малапропизмов может быть улучшено на основе более простого алгоритма, который анализировал семантическое расстояние между всеми терминами текста, а не на основании отношения с одной лексической цепочкой.

Тем не менее, работа [76] оказала сильное влияние на попытки моделирования построения лексических цепочек и применения их в разных компьютерных приложениях при автоматической обработке связного текста.

Описанный в этом разделе алгоритм является так называемым «жадным» (greedy) алгоритмом построения лексических цепочек, поскольку построение цепочек базируется только на словах, которые встречались ранее текущего кандидата. Такой алгоритм может образовать ложные цепочки из-за многозначности слов. Поэтому предложены также и нежадные алгоритмы построения лексических цепочек, которые предполагают построение полной картины возможных лексических отношений между кандидатами, предварительное разрешение лексической многозначности и только после этого построение лексических цепочек.

Примером нежадного алгоритма является подход к построению лексических цепочек, описанный в работе [194]. Алгоритм сначала выбирает существительные-кандидаты для построения лексических цепочек. На втором этапе устанавливаются все возможные отношения между всеми значениями кандидатов. В данном алгоритме рассматриваются такие отношения, как повторы, синонимы, гипонимы, гиперонимы, меронимы, холонимы и антонимы, также используются пути гиперонимических отношений, для которых длина пути не ограничивается. После установления всех возможных связей между словами, порождаются лексические кластеры. Лексические кластеры в данном алгоритме не являются взаимно исключающими, т.е. одно и то же слово может относиться к разным лексическим кластерам. На следующем шаге объединяются все лексические кластеры, относящиеся к одним и тем же значениям слов. Это дает возможность установления транзитивных отношений между словами, которые явным образом не указаны в WordNet.

Полученные лексические кластеры разбиваются на лексические цепочки так, чтобы между соседними элементами цепочки было не более 80 слов, и каждая цепочка состояла не менее чем из 3 слов. Эти цепочкам затем присваивается вес в зависимости от доли текста, которую занимает цепочка (фрагмент цепочки), и плотности цепочки (количество элементов цепочки по отношению к длине фрагмента цепочки).

Данный подход применялся к экспериментам по поиску документов по запросам конференции TREC и сравнивался с результатами работы известной информационно-поисковой системой, построенной на векторной модели, SMART [177]. Эксперименты показали, что система Stairmand находит релевантные документы лучше, если слова запроса относятся к основной теме или важной подтеме документа. Однако система SMART лучше различает между документами, которые частично относятся к теме запроса и нерелевантными документами. Кроме того, полнота работы алгоритма была очень низкой. Автор объясняет данную проблему

недостаточным покрытием WordNet реальных текстов, и особенно недостаточным описанием собственных имен в WordNet.

Рассматривая методы построения лексических цепочек с использованием лексических отношений, описанных в WordNet, авторы работы [14] указывают на проблему неправильного построения лексических цепочек за счет того, что выбор значений многозначных слов только на основе информации о предшествующих лексических цепочек не является достаточно качественным.

Поэтому в вышеупомянутой работе предлагается выделять все значения слов текста и встраивать их в начатые лексические цепочки. Понятно, что число вариантов цепочек даже для небольшого текста становится слишком большим. Чтобы снизить число вариантов в процессе обработки текста для каждой начатой цепочки оценивается ее сила, и в тот момент, когда количество вариантов превышает некоторый порог, удаляются наиболее слабые варианты цепочек.

Вес лексической цепочки определяется числом элементов цепочки и весом отношений между элементами цепочки. По завершении обработки текста наилучшая цепочка определяется как имеющая наибольшее число ребер графа цепочки (отношений между элементами цепочки) (см. рис. 3.2).

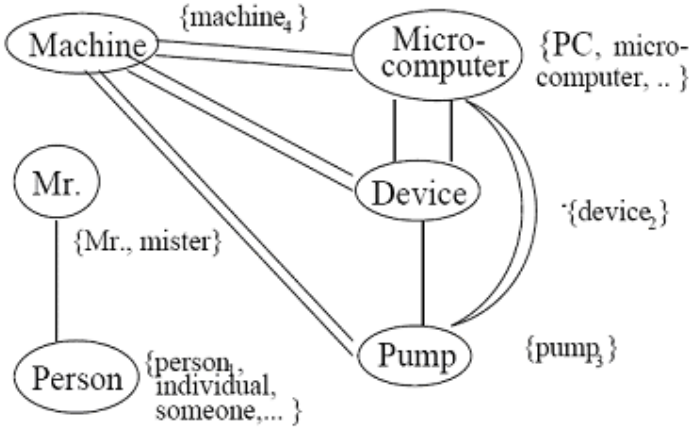


Рис. 3.2. Выбор значений слов на основе графа отношений в работе [14]

В данной работе было проведено исследование, на основе каких параметров выделенных лексических цепочек, можно отделить более сильные лексические цепочки. Было выявлено, что наилучшими показателями силы цепочки являются такие показатели, как длина цепочки Length, равная числу словоупотреблений в цепочке, и индекс гомогенности HomogeneityIndex, вычисляемый следующим образом:

$$HomogeneityIndex=1- (число\ разных\ слов\ в\ цепочке)/Length$$

Авторы работы, поэкспериментировав с разными формулами вычисления силы цепочки, остановились на следующей формуле:

$$Score\ (Chain) = HomogeneityIndex \cdot Length$$

Таким образом, вес цепочки фактически равен числу повторных употреблений слов в этой цепочке, и тем самым имеет прямую аналогию с частотой употребления слова в тексте. Снижение веса для цепочек со слишком разнообразным составом, видимо, позволяет снизить ошибки формирования лексических цепочек.

Для получения статуса сильной цепочки, которая будет использоваться в дальнейшем анализе, необходимо, чтобы для веса цепочки выполнялось следующее соотношение:

$$Score\ (Chain) > Average\ (Scores) + 2\ StandardDeviation\ (Scores)$$

Попытка тестирования качества таких лексических цепочек была выполнена в работе [183]. Предлагаемый метод тестирования основан на использовании аннотаций, созданных людьми.

Предполагается, что если лексические цепочки являются хорошим промежуточным представлением для отражения содержания документа, то можно ожидать, что существительные в таких аннотациях используются в том же самом смысле, что и существительные, сгруппированные в сильные лексические цепочки. Более того, сильные цепочки должны быть достаточно хорошо представлены в ручных аннотациях.

Для оценки использовался корпус из 10 научных статей, которые снабжены авторской аннотацией, а также 14 глав из 10 университетских учебников, для которых также имеются аннотации. Для каждого документа в корпусе, документ и его аннотация анализировались отдельно, и для каждого из них была построены лексические цепочки. Синсеты (значения) существительных в каждой из цепочек в документе и аннотации были сопоставлены между собой. При тестировании авторы вышеупомянутой статьи получили следующие результаты: 79.12% – существительных из сильных цепочек в документе содержатся в аннотации, 80.83% – существительных из сильных цепочек аннотации содержатся в документе.

Многие исследователи, исследующие лексическую связность на базе WordNet, отмечали, что серьезной проблемой является недостаточность лексических знаний, описанных в WordNet. В работах [197, 198] сделаны усилия для того, чтобы преодолеть эту проблему.

В данных работах предлагается дополнительно использовать следующую информацию:

- статистические ассоциативные связи слов,
- лексические цепочки для собственных имен.

Авторы этих работ подчеркивают, что одним из важных назначений учета статистических ассоциаций слов является преодоление уже упоминавшейся теннисной проблемы, т.е. проблемы, что в WordNet, слова, относящиеся к одной и той же тематической области, могут располагаться достаточно далеко по иерархии путей. Также авторы отмечают проблему нехватки такой информации, как некоторых значений, а также многословных сочетаний.

Для построения ассоциаций слов авторы использовали текстовый корпус конференции TDT (http://projects.ldc.upenn.edu/TDT/), извлекли из него все существительные и словосочетания WordNet и собрали информацию о совместной встречаемости существительных в пределах текстового окна, состоящего из четырех существительных. Окно было также ограничено

границами предложения и документа. Ассоциативные связи между словами, полученные на основе статистических критериев считаются самым слабым видом отношений между словами и применяются, если более сильных связей не найдено.

О. Медельян [136] предлагает использовать недостающее в WordNet ситуативное знание на основе информационно-поискового тезауруса (в работе используется тезаурус AGROVOC). Она указывает, что наиболее известные алгоритмы построения лексических цепочек слишком зависят от порядка слов в тексте, что не соответствует реальной ситуации, когда одно и то же содержание может быть выражено с помощью по-разному упорядоченных последовательностей предложений. Поэтому в работе предлагается сначала собрать цепочки-кандидаты со всего текста, а затем, получив целостную картину лексических цепочек-кандидатов текста, применить разбиение получившегося графа на наиболее связанные фрагменты.

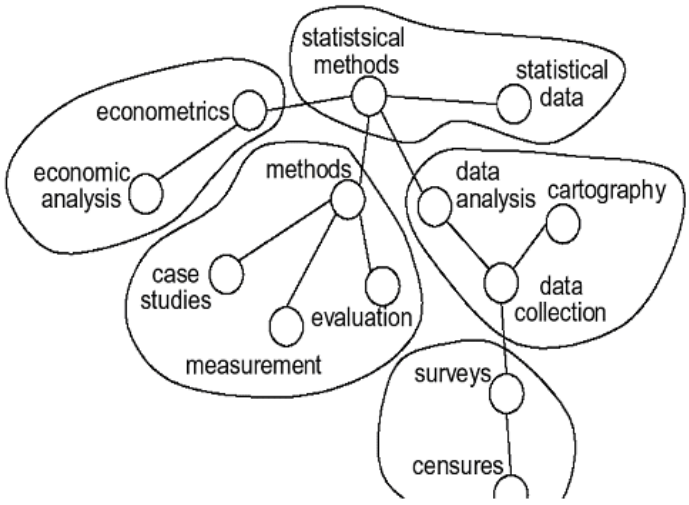


Рис. 3.3. Разбиение графа на лексические цепочки в работе [136]

Лексическая цепочка определяется как граф $G = (V, E)$ с узлами $v_i \in V$, представляющими термины тезауруса и дугами графа $(v_i, v_j, w_{ij}) \in E$,

описывающими отношения между терминами, где w_{ij} – это вес, выражающий силу отношения между терминами.

Такой граф строится следующим образом. Как и в предшествующих алгоритмах, цепочки-кандидаты строятся по порядку движения текста. Различия возникают в том случае, когда очередной термин может быть отнесен более чем к одной лексической цепочке. Тогда эти цепочки склеиваются в единую цепочку, а составные части этой единой цепочки удаляются из списка цепочек.

Получается граф достаточно сложной формы (см. рис. 3.3). Этот граф с помощью алгоритмов кластеризации графа разбивается на фрагменты так, чтобы между каждым элементом подграфа было расстояние не более 3 шагов, тем самым получаются сильно связанные между собой подграфы, которые и предлагается считать лексическими цепочками.

3.3. Автоматическое аннотирование текстов

Автоматически выявляемые лексические цепочки используются при решении разнообразных прикладных задач [205]:

- автоматической сегментации текстов [140, 141];
- автоматического разрешения многозначности [52];
- информационный поиск [194];
- автоматическое аннотирование текстов [14, 21, 182, 199];
- распознавание тем текстов [30],
- построение вопросно-ответных систем [142] и др.

Одним из самых популярных применений лексических цепочек является автоматическое аннотирование текстов, т.е. автоматическое порождение краткого изложения исходного текста [1]. Это связано с тем, что автоматическая аннотация должна быть понятной и связной, т.е. при ее создании полезно учитывать законы построения связных текстов.

Наиболее распространенном в настоящее время методом автоматического аннотирования текстов является составление аннотации из предложений (или фрагментов предложений) исходного текста, т.е. если текст T состоит из множества предложений $\{s_1 \ldots s_i \ldots s_n\}$, то множество предложений аннотации $A=\{s_1 \ldots s_k\}$ является подмножеством предложений текста: $A \subset T,\ k \leq n$

Существуют разные виды аннотаций [168]. Индикативная аннотация должна передать информацию об общем содержании документа, не сообщая деталей. Информативная аннотация должна сохранить информационную ценность исходного сообщения. Тематически-ориентированные аннотации должны отразить информацию из текста, соответствующую теме, интересующей пользователя, так называемые аннотации по запросу (query-based summaries). Экстрактивная аннотация состоит из фрагментов (предложений) исходного текста, в то время как аннотации в форме абстракта порождаются на основе извлеченного содержания.

Несмотря на существование ряда исследований по созданию аннотаций-абстрактов, основные исследования в настоящее время сосредоточены в сфере экстрактивных аннотаций.

Большинство систем аннотирования используют предложения исходного текста в качестве единиц порождаемой аннотации. Для предложения на основе выделенных характеристик подсчитываются веса, из предложений с наибольшими весами формируются аннотации.

Характеристики, на основе которых может составляться вес предложения, могут быть следующими:

- позиция в тексте,
- частотность слов,
- наличие ключевых фраз вида «Необходимо подчеркнуть»,
- длина предложения,
- именованные сущности,
- повторяемость слов и др.,

Современные подходы используют методы машинного обучения для учета возможных характеристик предложений, включаемых в аннотации [101].

Одним из относительно новых направлений составления аннотаций является составление аннотации на основе многих документов – обзорного реферата [41]. При составлении такого обзорного реферата необходимо решать такие задачи, как:

- борьба с избыточностью информации,
- идентификация важных различий между документами,
- обеспечение тематической связности текста, что усложняется тем, что предложения могут браться из разных источников.

Обзорные рефераты могут делаться для различных наборов документов [149], например, таких, как документы, описывающие конкретное событие, документы, обсуждающие одну и ту же тему, документы, обсуждающие биографию одного и того же человека, документы, обсуждающие множество событий одного и того же типа, например, конкретные примеры насилия, документы, представляющие мнения разных сторон на общую тему (например, мнение сената, конгресса, общественности на тему миграции).

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация – выделение близких по содержанию кластеров предложений [168]. Другим подходом к оценке избыточности является сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и оценка новой (непохожей) информации, например, так называемый подход Maximal Marginal Relevance (MMR) [28].

Обеспечение связности изложения является сложной проблемой, поскольку требует реального понимания содержания фрагментов и знаний о структуре связного текста. Многие подходы ограничиваются учетом времени

и порядка предложений в тексте (фрагмента из более раннего текста размещаются сначала, в порядке следования в тексте).

Оценка качества автоматически порождаемых аннотаций является сложной процедурой, поскольку даже для относительно содержательно простых документов как новостные сообщения, согласие между экспертами может составлять всего 60%.

Оценка качества аннотации может быть внутренней и внешней. Внутренняя (intrinsic) оценка аннотаций связана с оценкой качества аннотации как собственно текста, сравнения ее с исходным текстом или с аннотациями, порожденными людьми. При оценке качества аннотации экспертам могут быть заданы такие вопросы с оценкой по 5 бальной шкале:

- является ли предложения аннотации грамматически правильными,
- является ли текст аннотации связным,
- содержит ли аннотация все основные обсуждаемые темы исходного документа (документов) и др.

При оценке аннотаций по многим документам – обзорных рефератов в рамках конференции DUC, эксперты помимо ответа на конкретные вопросы по качеству аннотаций должны проставить и две общие оценки аннотации [41].

Во-первых, эксперты должны были оценить соответствие содержанию кластера, т.е. насколько реферат отображает необходимую для пользователя, формировавшего запрос, информацию. При этом не бралась в расчет читабельность реферата, до тех пор, пока она не влияла на объем покрытой в реферате информации.

Во-вторых, эксперты ставили общую оценку аннотации, которая должна отражать как содержательную часть реферата, так и его читабельность. При определении уровня общего соответствия оценщикам не предоставляли доступ к ранее оцененным характеристикам читабельности и соответствия содержанию, вместо этого они должны были «сходу» дать свою оценку. Многие из оценщиков посчитали для себя полезным выставлять

уровень общего соответствия исходя из ответа на вопрос: «Сколько я бы заплатил за этот обзорный реферат?». В итоге, плохая читабельность систем занижала их оценку общего соответствия, по сравнению с соответствием содержанию. В то же время, рефераты с высоким показателем читабельности, получали оценки за общее соответствие выше, по сравнению с оценками за соответствие содержания.

Внешняя (extrinsic) оценка аннотации производится в специально поставленной задаче, в которой выясняется, может ли аннотация заменить исходный текст. Такими задачами могут быть классификация документов по его аннотации, или ответы на вопросы по содержанию документа на основе его аннотации.

Один из первых масштабных экспериментов по внешней оценке аннотаций был осуществлен в рамках конференции SUMMAC [127]. В оценку было включено три задачи:

- задача классификации (насколько качество классификации документа по аннотации сравнимо с качеством классификации полного документа),
- ad hoc задача – эксперты должны определить, насколько текст соответствует запросу по аннотации,
- вопросно-ответная задача – эксперты должны ответить на вопросы по основному содержанию документа на основании его аннотации.

Важным элементом современной оценки аннотаций является получение автоматических оценок качества аннотаций за счет автоматического сравнения порожденной аннотации с аннотациями, написанными людьми.

В рамках конференции DUC используется метод автоматической оценки качества аннотаций ROUGE (Recall Oriented Understudy for Gisting Evaluation), который подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми [104].

Применение лексических цепочек для автоматического аннотирования позволяет решать несколько задач, возникающих в процессе автоматического аннотирования документов. Они помогают выявлять основную тему документа, и, кроме того, являются дополнительным фактором обеспечения связности создаваемой аннотации [170, 233, 277, 292]. Рассмотрим подробнее некоторые из предлагаемых подходов по использованию лексических цепочек для порождения разного вида аннотаций.

Одной из первых работ, описывающих применение алгоритмов выявления лексических цепочек, к автоматическому аннотированию текстов, была работа [14]. Как указывалось выше, в этой работе был реализован алгоритм построения лексических цепочек на основе WordNet, а также были сделаны усилия, чтобы разобраться, какими свойствами должны обладать так называемые сильные лексические цепочки, т.е. цепочки, которые наилучшим образом отражают содержание текста.

Идея применения лексических цепочек для автоматического аннотирования документов состоит в том, что если цепочка отражает важные темы документа, то необходимо для аннотации выбирать предложения, в которых встречались элементы этих важных цепочек. Конкретный алгоритм был следующим: для каждой цепочки выбирается ее представители – элементы цепочки, частотность которых превышает среднюю частотность элементов цепочки. Для составления аннотации берутся первые по порядку текста предложения, которые содержат элемент-представитель для каждой из сильных лексических цепочек. Таким образом, каждая сильная лексическая цепочка представлена, по крайней мере, одним предложением в аннотации.

Для оценки качества предложенного метода автоматического аннотирования было выбрано 40 новостных текстов, каждый в среднем по 30 предложений. Пять ассессоров должны были сделать два вида аннотаций для этих текстов длиной 10% и 20% от длины исходного текста.

На основе этих аннотаций была сформирована «идеальная» аннотация, которая содержала те предложения, которые были выбраны большинством

ассессоров. Автоматически порождаемые аннотации были сравнены с аннотациями, порожденный суммаризатором Microsoft Word (табл. 3.1), посредством вычисления показателей полноты и точности:

	Microsoft		Лексические цепочки	
	Точность	Полнота	Точность	Полнота
10%	33%	37%	61%	67%
20%	32%	39%	47%	64%

Таблица 3.1. Результаты сравнения аннотаций, построенных на основе лексических цепочек с суммаризатором Microsoft Word.

В таблице (3.1) видно, что аннотации, построенные на базе лексических цепочек, в значительной степени ближе к аннотациям, порождаемым людьми.

В работе [45] алгоритм автоматического аннотирования из работы [14] тестируется на основе внешней задачи, а именно в рамках задачи автоматического нахождения похожих текстов. Предполагается, что если автоматическая аннотация хорошо отражает основное содержание документа, то аннотации похожих документов будут также похожи, а аннотации разных документов также будут различаться.

Подход [14] сравнивался с тремя базовыми подходами: случайным выбором предложения, выбором блока первых предложений, выбором предложений на основе метрики tf.idf. Тестирование проводилось для разных коэффициентов сжатия от 10% аннотации до 60% аннотации. Подход [14] уступил базовым подходам только 1 раз: при 10% аннотации лучшими были аннотации, построенные на основе первых предложений исходных текстов.

В работе [21] аннотации строятся на основе другого рода лексических цепочек. Используется «жадный алгоритм» из работы [76], который имеет следующие дополнения:

- длина пути между элементами цепочки не более 2 отношений,
- такие отношения должны быть между всеми элементами цепочки.

Наиболее значительное отличие данного подхода от других подходов заключается в том, что делается дополнительный предварительный шаг по отбору существительных – кандидатов для включения в лексические цепочки. В большинстве подходов предварительная стадия построения лексических цепочек включает морфологический анализ и отбрасывание стоп-слов, которые часто дают ошибочные или малоинформативные лексические цепочки. В данной работе проверяется предположение о том, что существительные, находящиеся в подчинительных предложениях, менее информативны, и их можно не включать в процесс построения лексических цепочек.

В работе [100] исследуется возможность использования лексических цепочек для построения обзорного реферата по запросу. Построение лексических цепочек производится для получения наиболее сильных цепочек, в терминах работы [14]. Построение лексических цепочек в этой работе проводится в два этапа. На первом этапе строятся отдельные лексические цепочки, на втором этапе построенные лексические цепочки корректируются.

Построение цепочек происходит, начиная с самых частотных синсетов. В начатую лексическую цепочку вносятся все синсеты, которые могут быть отнесены к синсетам цепочки по принятой мере близости. Этот процесс проводится для наиболее частотной половины из всех синсетов-кандидатов, для которых могут быть построены лексические цепочки. После построения цепочек определяются наиболее сильные цепочки. На втором этапе сильные цепочки, содержащие хотя бы одно общее слово, сливаются в единую лексическую цепочку.

Для порождения аннотации по запросу из набора документов извлекаются предложения, имеющие наиболее высокий вес по следующей формуле:

$$Score = \alpha P(chain) + \beta P(queries) + \gamma P(nameentity),$$

где $P(chain)$ – это сумма весов лексических цепочек, участники которых были упомянуты в предложениях-кандидатах, $P(queries)$ – это сумма совпадающих слов в предложении-кандидате и формулировке темы запроса, $P(nameentity)$ – это число именованных сущностей, упомянутых как в предложении-кандитате, так и формулировке запроса.

Таким образом, исследователи связного текста выделяют несколько взаимосвязанных между собой видов связности текста. Среди всех видов связности лексическая связность наилучшим образом поддается моделированию на основе информации, описанной в тезаурусах и онтологиях.

При моделировании лексической связности существенным является не установление пар лексически связанных слов, а цепочек близких по смыслу слов, так называемых лексических цепочек. Получение таких лексических цепочек важно не само по себе, а как шаг к выявлению тематической структуры текста, т.е. определению основной темы и побочных тем (подтем) документа.

Алгоритмы, основанные на лексических цепочках, использовались при решении различных задач автоматической обработки текстов. Особенно популярны методы, основанные на лексических цепочках, в задаче автоматического порождения аннотаций для одного и многих документов, поскольку именно в этой задаче особенно важно обеспечить связность порождаемой аннотации. Также лексические цепочки в автоматическом аннотировании помогают снизить излишние повторы в порождаемых аннотациях.

3.4. Проблемы автоматического построения лексических цепочек

Как уже указывалось, описания языковых выражений в тезаурусах могут использоваться для выявления лексической связности текста, что

обычно делается посредством построения так называемых лексических цепочек – совокупностей языковых выражений текста, близких по смыслу.

Основными критериями для построения лексических цепочек в большинстве подходов являются следующие:

- наличие и сила связей между лексемами, описанных в некотором ресурсе,
- расстояние между вхождениями лексем в тексте, измеряемое обычно в предложениях. Если расстояние от текущего слова до предшествующих вхождений лексической цепочки больше некоторого порога, то лексическая цепочка прерывается и начинается новая.

Возникает вопрос, достаточно ли вышеперечисленных критериев для построения лексических цепочек.

Второй вопрос, возможно связанный с первым, заключается в том, что являются ли лексические цепочки такими уж очевидными, поскольку, как мы увидим ниже, эксперименты по сравнению лексических цепочек, выделенных разными людьми, показали достаточно серьезное расхождение в представленных лексических цепочках. Второй вопрос связан с первым, так как важно понять, является ли такая субъективность неизбежной, или не учитывается какой-либо важный критерий построения лексических цепочек.

3.4.1. Субъективность выделения лексических цепочек

Авторы работы [78] указывают на субъективность рассмотрения лексической связности в тексте. Они рассматривают пример небольшого текста:

> *(*) How can we figure out what a text means. One could argue that the meaning is in the mind of the reader, but some people think that the meaning lies within the text itself."*

Отвечая на вопрос, каковы лексические цепочки, которые можно выделить в данном тексте, один автор статьи полагает, что видит две цепочки: «понимание», которые включают такие слова, как *figure out, means, meaning, mind, think, meaning* и цепочка «текст», включающая слова *text, reader, text.* Второй автор также выделил две цепочки, но соотнес слова *means, meaning* с цепочкой «текст».

Действительно, при построении лексических цепочек текста (*) слова *значение, значить* близки по смыслу как слову *текст*, так и словам *думать, узнать*. Можно ли определить, кто из авторов статьи прав, или, может быть, слова «значение» и «значить» входят в две лексические цепочки?

Также в [78] описывается эксперимент по изучению согласия между читателями по выявлению лексической связности текста, в котором показано, что при выделении в тексте связанных по смыслу групп слов коэффициент согласия составил 63%.

В работе [80] описывается эксперимент по сравнению лексических цепочек, создаваемых разными людьми, на примере научной статьи Lee Lilian “Measures of distributional similarity”, опубликованной в трудах 37 конференции ACL (pp. 25-32). В эксперименте участвовали 3 человека, которым было дано неограниченное время, чтобы создать наборы терминов, которые им кажутся близкими по смыслу в контексте исследуемой статьи. Каждому аннотатору были даны список всех слов статьи, упорядоченные по мере частотности и максимальные именные группы, извлеченные из текста. Использование этих материалов носило вспомогательный характер.

В статье [80] приводятся лексические цепочки, полученные двумя аннотаторами. В каждой цепочке выделен наиболее частотный элемент, который является как бы представителем цепочки. Один аннотатор создал 12 лексических цепочек, второй аннотатор создал 22 лексические цепочки, причем имеется совпадение главных элементов лексических цепочек только в четырех случаях (с точностью до единственного/множественного числа): *similarity, probability, cooccurrance, distribution.*

Таким образом, в экспериментах были выявлены значительные расхождения в формировании лексических цепочек людьми, и возникает вопрос, является ли эта ситуация стандартным проявлением субъективности человеческих решений или при рассмотрении лексических цепочек не учитываются какие-то дополнительные факторы.

3.4.2. Построение лексических цепочек с учетом ситуативных отношений

Стандартным базовым ресурсом для построения лексических цепочек является тезаурус WordNet. Однако набор отношений в этом тезаурусе невелик. Многие авторы, занимавшиеся автоматическим построением лексических цепочек, указывали на одну из проблем построения лексических цепочек по WordNet – нехватку ситуативных отношений (см. п. 3.2). Но появление такого рода отношений в ресурсе, опять ставит вопрос о критериях выделения цепочек.

Рассмотрим следующий текст на медицинскую тему:

(**)

*Канадские **врачи** убили **пациента** передозировкой **наркотика***

*В Канаде начато расследование **несчастного случая** в **больнице** города Ред Дир, где **медики** по ошибке ввели **пациенту смертельную** дозу **опиоидного наркотика**, сообщает газета The Globe amd Mail. 69-летний **пациент** поступил в **приемное отделение больницы** после **травмы грудной клетки**, которую он получил во время конной прогулки. **Врач** назначил ему 10 миллиграммов **морфина** в качестве **обезболивающего** и отпустил домой.*

*По ошибке **медсестры пациенту** был сделан укол **гидроморфона** – похожего на **морфин** по названию и действию. Однако этот **препарат** гораздо сильнее – доза в 10 миллиграммов **смертельна**. Свою ошибку **медики** осознали после пересчета **наркотических средств** и сразу позвонили родственникам мужчины. Однако состояние **пациента** быстро ухудшилось, и он **умер** после возвращения в **больницу**.*

Расследование этого случая завершится в течение 10 дней. Как сообщают в ***больнице, укол*** *сделала опытная* ***медсестра****, которая полностью признает свою ошибку. Однако есть вероятность, что после расследования ее все же признают невиновной. По заявлению министра* ***здравоохранения*** *провинции Альберта, главное – сделать, чтобы такая ошибка не повторилась. (*Источник: Mednovosti.ru)

В тексте содержится множество слов и словосочетаний, имеющих отношение к медицине: *наркотики, больница, пациент, травма, морфин, обезболивающее, гидроморфон, медик, врач* и др. Возникает вопрос, должны ли все эти слова собраться в одну лексическую цепочку или несколько. Если разбивать на несколько лексических цепочек, то нужно понять, какие формальные критерии должны быть применены.

Следствием более богатой системы отношений в лингвистическом ресурсе является и то, что одно и то же слово может быть отнесено к разным лексическим цепочкам, хотя, как уже указывалось, основополагающим принципом подавляющего большинства подходов, в которых изучается автоматическое построение лексических цепочек, является отнесение очередного слова только к одной лексической цепочке. Например, в тексте про Украину словосочетание *Верховная рада* может быть в равной степени отнесена к двум лексическим цепочкам – цепочке парламента *(парламентских выборов, депутаты, депутатов, Верховную Раду, парламента) и* цепочке Украины *(президент Украины, украинского, Верховную Раду, Украине*).

То, что в реальной ситуации одно и то же слово может быть отнесено к разным цепочкам одновременно, значительно усложняет алгоритмы автоматического построения лексических цепочек.

Мы нашли только одну работу [80], в которой авторы указывают на то, что их алгоритм построения лексических цепочек позволяет относить одно и то же слово или словосочетание к разным лексическим цепочкам, и при этом

они указывают на проблему порождения слишком большого количества лишних лексических цепочек (overgeneration). При этом авторы подчеркивают, что в проведенном ими эксперименте все эксперты-аннотаторы, по крайней мере, одно слово (словосочетание) отнесли более чем к одной лексической цепочке.

3.5. Модель тематического представления текста

3.5.1. Лексические цепочки и тематическая структура текста

Во всех подходах автоматического моделирования лексических цепочек построение этих цепочек не является самоцелью – лексические цепочки выделяются для того, чтобы «приблизиться» к автоматическому построению тематической структуры текста, т.е. уметь выделять, что в тексте главное, что второстепенное, как текстовые сущности связаны друг с другом.

С целью выделения наиболее значимых для содержания текста лексических цепочек, рассматриваются различные параметры лексических цепочек, такие, как частотность ее элементов, текстовое покрытие и другие. В лексических цепочках выделяются наиболее частотные элементы цепочки в качестве наиболее важных тематических элементов текста.

Поскольку целью автоматического выделения лексических цепочек является автоматическое построение тематической структуры текста, рассмотрим на методы построения лексических цепочек и вышеописанные проблемы их построения с точки зрения роли лексических цепочек в тематической структуре текста.

Многие исследователи указывают на то, глобальная связность текста проявляется в том, что текст имеет единую тему. Тематическая структура текста представляет собой иерархическую структуру тем и подтем. Каждому предложению текста имеется некоторое соответствие в этой тематической структуре. Таким образом, предполагается, что содержание текста выражается в виде совокупности пропозиций: $P^D = \{p_0 (c_{01}...c_{01}), p_1 (c_{11}...c_{1n})..$ $p_k (c_{k1}...c_{kn})\}$, где c_{ij-} это понятия или экземпляры, упомянутые в тексте D. Над

этим множеством определено отношение частичного порядка, т.е. выполняются следующие свойства:

- рефлексивность,
- транзитивность,
- антисимметричность

У связного текста имеется основная тема – главная пропозиция p_0 ($c_{01}…c_{0n}$) (макропропозиция по терминологии Ван Дейка [42]):

$$\forall p_i ((p_i \in P^D) \rightarrow (p_i \leq p_0))$$

Аргументы пропозиций $c_{i1}…c_{in}$ будем называть тематическими элементами, а аргументы основной темы документа $c_{01}…c_{0n}$ - основными тематическими элементами документа. По своей природе тематические элементы представляют собой понятия или идентификаторы конкретных объектов.

Основная пропозиция p_0 ($c_{01}…c_{0n}$) обычно представляет собой следующие частные случаи:

- p_0 (c_{01})- пропозицию над одним атрибутом – например, описание компании, или биография человека;
- p_0 ($c_{01}…c_{0n}$) = r_s ($c_{01}…c_{0n}$) – описывает взаимоотношения между тематическими элементами $c_{01}…c_{0n}$ в некоторой ситуации;
- p_0 ($c_{01}…c_{0n}$) = r_{ss} ($r_{s1,}$ r_{s2}, $c_{01}…c_{0n}$) – описывает отношения между двумя ситуациями.

Пропозиции тем (подтем) устанавливают отношения между тематическими элементами $c_1…c_n$. В иерархической тематической структуре главная тема p_0 ($c_{01}…c_{0n}$) поясняется, характеризуется, дополняется деталями посредством подтем p_1 (c_{11}, ….c_{im}) … p_i (c_{i1}, …c_{ij}…c_{ik}).

Что представляют собой тематические элементы подтем c_{ij} по отношению к тематическим элементам основной темы текста c_{0k}?

Рассмотрим две пропозиции p_i (c_{i1}, c_{i2} … c_{ik}) и p_j (c_{j1}, c_{j2} … c_{jm}) такие, что $p_i \in P^D$, $p_j \in P^D$, $p_i \leq p_j$.

Такие пропозиции связаны между собой и поэтому должны существовать взаимоотношения между участниками этих пропозиций, т.е.

$$\forall p_i, p_j \; ((p_i \leq p_j) \rightarrow \exists \; c_{il} \, c_{jn} \, r_c \, (c_{il}, \, c_{jn}))$$

По своей природе r_c может быть отношением кореферентности r_{ref}, т.е. c_{il} и c_{jn} являются ссылками на один и тот же объект действительности, и/или между c_{il} и c_{jn} существует известное лексическое отношение r_l, описанное в ЛО (точный повтор, синонимический повтор, родовидовые отношения, отношения часть-целое и др.). Таким образом, в силу глобальной связности текста в каждой подтеме, по крайней мере, один тематический элемент (а часто и больше) должен соответствовать тематическим элементам основной темы текста.

В результате каждый тематический элемент c_{0i} основной пропозиции p_0 имеет представительство в пропозициях нижнего уровня p_j посредством связанных с ним по смыслу элементов пропозии $l_{0ij.}$ Возникает структура типа узла: основной тематический элемент c_{0i} и связанные с ним элементы $l_{0ij.}$ Мы называем такой узел тематическим узлом $tnode_i$:

$$tnode_i = (c_{0i,} \, \{l_{0i1} \ldots l_{0ij} \})$$

Множество всех тематических узлов, выделяемых в тексте, будем обозначать $Tnode = \{tnode_1 \ldots \; tnode_n\}$.

Таким образом, основная роль лексических цепочек относительно тематической структуры текста состоит в обеспечении представительства тематических элементов более высоких уровней иерархии в подтемах более низкого уровня (см. рис. 3.4).

Отсюда следует, что в «правильной» совокупности лексических цепочек текста, т.е. в лексических цепочках, отражающих тематическую структуру анализируемого текста, каждому тематическому элементу основной темы текста должны соответствовать свои лексические цепочки (которые могут иметь пересечение в некоторых словах).

Кроме того, лексические цепочки действительно имеют наиболее важных представителей – это элемент темы более высокого уровня. Рядовые элементы цепочки – это тематические элементы нижестоящих тем, раскрывающих эту тему.

Таким образом, на наш взгляд, по внутренней структуре лексическая цепочка имеет структуру узла с выделенным центральный элементом и некоторой совокупностью лексем, связанных с этим центральным элементом. Назовем лексическую цепочку с такой предполагаемой структурой *тематическими узлом*.

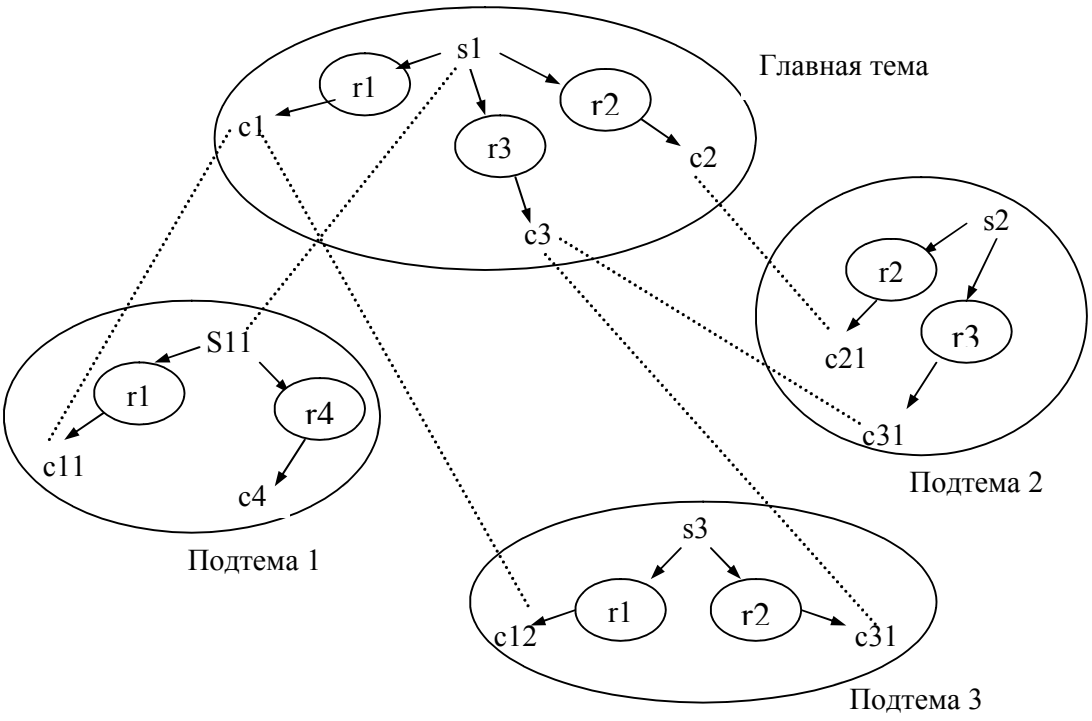


Рис. 3.4. Тематическая структура текста как иерархия пропозиций тем: s_i – ситуации, описываемые в предложениях текста, c_j – понятия, участвующие в ситуциях, r_k – роли понятий в ситуациях

Среди тематических узлов можно выделить основные тематические узлы и локальные тематические узлы. Основные тематические узлы имеют в качестве центра тематические элементы основной темы документа.

Пропозиция основной темы документа, т.е. взаимоотношения участников основной темы, также должна находить свое отражение в конкретных предложениях текста, которые должны раскрывать, уточнять взаимоотношения между тематическими элементами. Если текст посвящен обсуждению взаимоотношений между тематическими элементами $c_1 \ldots c_n$, то в предложениях текста должны обсуждаться детали этих отношений, что проявляется в том, что сами тематические элементы $c_1 \ldots c_n$ или их лексические представители должны встречаться как разные актанты одних и тех же предикатов в конкретных предложениях текста.

Отсюда следует практический вывод: если даже очень близкие по смыслу лексические сущности c_1 и c_2 часто встречаются в анализируемом тексте в одних и тех же простых предложениях, то это означает, что данный текст посвящен рассмотрению отношений между этими сущностями, т.е. c_1 и c_2 соответствуют разным тематическим элементам основной темы или подтемы текста и должны быть отнесены к разным лексическим цепочкам (тематическим узлам).

Таким образом, «правильные» лексические цепочки, отражающие тематическое содержание документа должны отвечать следующим условиям:

1) лексическая цепочка имеет внутреннюю структуру узла – к одному выделенному элементу относятся все другие элементы лексической цепочки;
2) лексическая цепочка не должна содержать слова и словосочетания, которые часто встречались в одних и тех же предложениях текста с главным элементом этой цепочки, поскольку частая встречаемость некоторой лексической единицы l_i с начальным элементом цепочки l_0 может означать, что l_i и l_0 представляют собой равноправные элементы основной или локальной темы анализируемого текста;
3) значимость цепочки для отражения содержания текста определяется не столько длиной, покрытием и другими характеристиками цепочки, а тем, насколько часто элементы этой цепочки встречались

с элементами других цепочек в одних и тех же предложениях текста, т.е. насколько много пропозиций конкретных предложений текста было посвящено обсуждению отношений между элементами некоторой совокупности лексических цепочек.

Для проверки вышеуказанной мысли о том, что близкие по смыслу слова чаще встречаются в соседних предложениях, чем в одних и тех же простых предложениях текста, был предпринят следующий эксперимент. Десять новостных кластеров, каждый с количеством документов более 40, были сопоставлены с Общественно-политическим тезаурусом [283], и были выделены пары наиболее близких по смыслу слов (выражений): синонимы, слова-виды, слова-типы и др.

Новостной кластер не является единым связным текстом, но тексты кластера посвящены одной теме, и поэтому статистическое проявление свойств связного текста в новостном кластере многократно усиливается, что мы и пытаемся использовать для эксперимента.

Таким образом, исследовалось совместное употребление следующих типов пар слов:

- синонимы-существительные (*Киргизия – Киргизстан*),
- дериваты прилагательные-существительные (*Киргизия – Киргизский*),
- отношения род-вид существительных (*депутат – представитель*),
- отношение род-вид прилагательное-существительное (*национальный – Россия*),
- отношение часть-целое существительные (*парламентарий – парламент*),
- отношение часть-целое прилагательное-существительное (*американский – Вашингтон*),
- отношение ассоциации существительные (*нефть – баррель*),
- отношение ассоциации существительное-прилагательное (*нефтяной – баррель*).

Для всех таких типов пар слов, употреблявшихся в текстах кластера с частотностью более четверти числа документов кластера, было подсчитано отношение встречаемости в пределах сегментов без запятых F_{segm} и встречаемости в соседних предложениях F_{sent}. В табл. 3.1 отражены результаты данного подсчета.

Табл. 3.1. Соотношение частотностей встречаемости близких по смыслу пар выражений внутри сегментов предложений и в соседних предложениях F_{segm}/F_{sent}

Тип отношения	F_{segm}/F_{sent}	Количество пар
синонимы-существительные	0.309	31
дериваты прилагательные-существительные	0.491	53
Отношение род-вид между существительными	1.130	88
отношение род-вид прилагательное-существительное	1.471	28
отношение часть-целое существительные	0.779	58
отношение часть-целое прилагательное-существительное	1.580	29
отношение ассоциации существительные	1.248	37
отношение ассоциации существительное-прилагательное	0.898	18
выражения, между которыми не установлены перечисленные типы отношений	1.440	21483

3.5.2. Примеры разбора лексических цепочек с учетом тематической структуры текста

Рассмотрим, каким образом выводы предыдущего раздела могут уточнить процедуру выделения лексических цепочек в текстах (*) и (**) из п. 3.4.1.

При анализе текста (*) возник вопрос, куда отнести слова *means, meaning* к цепочке *figure out, think…*, или к цепочке *text, reader*. Учитывая сделанные выводы, можно заметить, что в таком маленьком тексте слова *means, meaning* трижды встретились в одних и тех же простых предложениях со словами *text, reader:*

*what **a text means***
*the **meaning** is in the mind of the **reader***
*the **meaning** lies within the **text** itself*

Это означает, что данный текст посвящен рассмотрению отношения *текст – значение. Текст и значение* представляют собой разные тематические элементы в основной теме текста, и, соответственно, правильная структура лексических цепочек должна отнести слова *текст* и *значение* к разным лексическим цепочкам.

В то же время слова *means, meaning* не стоит относить и к другой лексической цепочке *figure out, think*, поскольку у этих глаголов один из актантов представляет собой клаузы, в которых и упоминаются слова *means, meaning,* т.е. опять же это является центральной темой фрагмента, что люди думают по поводу значения текста.

***figure out** what a text **means**…*
***think** that the **meaning** lies within the text itself.”*

Таким образом, лексические цепочки данного текста таковы:

1) text, reader, text.

2) figure out, think

3) means, meaning, meaning

В тексте (**) заголовок достаточно подробно называет основные тематические элементы текста*: врач (точнее медицинский работник), убить, пациент, наркотик*. И, действительно, мы видим повторяющуюся встречаемость этих тематических элементов в одних и тех же предложениях текста:

медики *по ошибке ввели* ***пациенту смертельную*** *дозу* ***опиоидного наркотика***

Врач *назначил ему* ***(пациенту)*** *10 миллиграммов* ***морфина****.*

По ошибке ***медсестры пациенту*** *был сделан укол* ***гидроморфона***

Свою ошибку ***медики*** *осознали после пересчета* ***наркотических средств***

Таким образом, в тексте (**) должны быть выделены, по крайней мере, три «медицинские» лексические цепочки:

- цепочка «медработники» (врачи, медики, врач, медсестры, медики, медсестра),
- цепочка «пациент» (пациент, пациенту, пациент, пациенту, пациента),
- цепочка «наркотик» (наркотика, наркотика, морфина, гидроморфона, морфин, препарат, наркотических средств).

Кроме того, отдельно может быть выделена лексическая цепочка «больница» *(больнице, приемное отделение, больницы, больницу, больнице*), элементы которой также встречаются в одних и тех же предложениях текста с представителями других медицинских цепочек:

в ***больнице*** *..., где* ***медики*** *по ошибке ввели смертельную дозу опиоидного* ***наркотика,***

пациент *поступил в* ***приемное отделение больницы,***

он (пациент) *умер после возвращения* ***в больницу,***

Как сообщают в ***больнице,*** *укол сделала опытная* ***медсестра***

Таким образом, анализ предложений текста позволяет выявить, что лучшим представлением для отражения содержания этого текста является не одна медицинская цепочка, а четыре цепочки, каждая из которых соответствует отдельному тематическому элементу данного текста, взаимодействующего с другими тематическими элементами.

Рассмотрение лексических цепочек через призму их употребления в одних и тех же предложениях текста имеет прямое соответствие с идеей Р. Хазан о «гармонии связности», которая проявляется в том, что элементы разных лексических цепочек должны выступать по отношению друг к другу в одних и тех же семантических отношениях, и, это значит, в большинстве случаев представители этих цепочек должны упоминаться в одних и тех же предложениях текста [73]. В одном из рассмотренных текстов – тексте (**) элементы четырех медицинских лексических цепочек четко находились по отношению друг к другу в одних и тех же семантических отношениях ‘агент’(медики)-‘пациенс’(пациент)-‘средство’(наркотик)- ‘место‘(больница).

Различие нашего подхода от идеи Р. Хазан заключается в следующих положениях. Во-первых, мы не требуем, чтобы непременно между элементами лексических цепочек были одни и те же семантические отношения, полагая, что уже частое упоминание элементов разных лексических цепочек в связном тексте не может быть случайным.

Во-вторых, рассмотрение синтагматических отношений между элементами потенциальных лексических цепочек является важным уже на этапе построения лексических цепочек. Это рассмотрение позволяет в сложных случаях употребления в тексте большого количества близких по смыслу слов принимать более обоснованное решение по разделению этого множества слов на лексические цепочки.

3.5.3 Автоматическое построение тематического представления

Мы предположили, что лексические цепочки должны связывать не все близкие по смыслу слова текста, но соответствовать тематической структуре текста. Кроме того, лексические цепочки должны иметь форму узла – с главным выделяемым элементов, к которому относятся все другие элементы этой цепочки. Далее таким образом устроенные лексические цепочки будем называть тематическими узлами Tnode.

Важно еще подчеркнуть, что поскольку тематические узлы призваны моделировать основное содержание текста, то тематические узлы – это не последовательности близких по смыслу лексем, а совокупности близких по смыслу понятий, т.е., сущностей в которых до какой-то степени устранен фактор лексической синонимии и многозначности.

В предыдущем разделе мы показали, что создать «правильный» (то есть соответствующий тематической структуре анализируемого текста) тематический узел невозможно, используя только локальную информацию о расположении слов в соседних предложениях документа. Нужна совокупная информация о частотности и распределении слов в тексте, которую необходимо сопоставить с имеющимися в ЛО знаниями о существующих соотношениях значений слов.

Поэтому лексические цепочки в форме тематических узлов не строятся при движении от предложения к предложению, а производятся из общей картины упоминания понятий в предложениях, полученной по тексту.

Для построения тематических узлов существенны два фактора:

- существование пути определенного вида между понятиями ЛО и
- встречаемость понятий ЛО в одних и тех же простых предложениях текста.

При изложении методов построения лексических цепочек на базе тезауруса WordNet используются некоторые типы путей между синсетами, в

том числе пути, состоящие из отношений различной направленности, т.е. пути с перегибами.

При построении тематических узлов на основе предлагаемой модели лингвистической онтологии ЛО мы отказались от использования путей с перегибами (P_{updown}, P_{downup} – см. п. 2.3) по следующим причинам:

Во-первых, в ЛО имеется больший набор прямых связей между понятиями за счет транзитивных отношений *часть-целое* и отношений направленной ассоциации, описывающих родовую зависимость понятий ЛО друг от друга. Во-вторых, мы считали важным дать возможность понятию ЛО входить в несколько тематических узлов. В-третьих, понятия, соединенные путями с перегибами – виды одного рода, части одного целого и др. – достаточно часто могут выступать как разные, противопоставленные друг другу элементы основной темы.

Таким образом, в основном блоке текущей реализации алгоритма тематические узлы образуются на основе иерархически подчиненных понятий ЛО, имеющих между собой пути, состоящие из отношений одной направленности – P_{up}, P_{down} (см. п. 2.3).

Для учета совместной встречаемости понятий ЛО в одних и тех же предложениях текста, для каждого понятия подсчитываются понятия-соседи в линейном контексте внутри предложения. Величина линейного контекста обычно устанавливается равной 3, т.е. для каждого понятия запоминается по три понятия-соседа влево и вправо. Понятия-соседи суммируются по всему тексту, и, таким образом, для каждого понятия получается частотный список понятий-соседей – так называемые текстовые связи понятия – r_{text} [275].

3.5.3.1. Алгоритм построения тематических узлов

Для автоматического построения тематических узлов *Tnode* сначала выделяются потенциальные центры тематических узлов. Предполагается, что то понятие ЛО, которое наиболее точно характеризует развиваемую в тексте тему и которое, соответственно, может стать тематическим центром одного

из тематических узлов текста, обычно некоторым образом выделяется в пространстве всех тематически близких понятий, а именно: такое понятие может быть упомянуто в заголовке и/или в начале текста, или имеет максимальную частотность среди других близких по смыслу понятий.

Тематическим центром может стать любое понятие ЛО, независимо от уровня его общности/специфичности. Единственное условие, которое может быть указано, это общая тематическая принадлежность концепта. При обработке современной прессы, актов законодательства на базе тезауруса РуТез обычно требуется принадлежность начального понятия тематического узла Общественно-политическому тезаурусу, т.е. фактически принадлежность понятия к одной из тематических областей общественной жизни.

Таким образом, создание тематического узла начинается с выбора главного понятия тематического узла. Сначала тематические узлы собираются вокруг понятий заголовка и первого предложения текста. Затем тематические узлы собираются для остальных понятий, начиная с самых частотных. Те понятия, которые уже попали в тематический узел некоторого понятия, свой тематический узел не образуют [275, 282].

Центральное понятие тематического узла c_0 присоединяет в создаваемый тематический узел понятия c_i из своей окрестности при выполнении нескольких условий. При присоединении учитываются такие факторы, как:

- количество текстовых связей между c_i и c_0 (т.е. совместной встречаемости c_i и c_0 в одних и тех же предложениях) в целом документе – r_{text},
- количество связей между c_i и c_0 по предложениям, т.е. сколько раз в документе c_i и c_0 встречались в текущем предложении и в k (по умолчанию k=7) соседних предложениях, но вне пределов окна установления текстовых связей – r_{sent}.

В новый тематический узел $Tnode_0$ понятия c_0 включаются понятия c_i из понятийной окрестности $O(c_0)$ при выполнении одного из следующих условий:

- $(r_{sent}(c_0, c_i) > 0 \wedge ((r_{text}(c_0, c_i) < 2) \vee (r_{text}(c_0, c_i) \leq r_{sent}(c_0, c_i)))$, т.е. понятия c_0 и c_i должны встречаться в тексте в соседних предложениях и при этом либо практически не встречаться рядом друг с другом в одних и тех же предложениях текста, либо частотность встречаемости понятия c_0 и c_i в одних и тех же предложениях текста должна быть меньше, чем частотность встречаемости в c_0 и c_i в соседних предложениях,

или

- $(r_{sent}(c_0, c_i) = 0 \wedge r_{text}(c_0,c_i) = 0 \wedge \exists\, c_t (c_t \in O(c_0) \wedge c_t \in Tnode_0 \wedge r_{sent}(c_t, c_i) > 0))$, т.е. понятие c_i из понятийной окрестности c_0 включается в тематический узел, если оно нашлось относительно недалеко от понятия c_t, уже включенного в тематический узел c_0.

или

- частотность упоминания понятия c_i равна 1.

После построения очередного тематического узла выбирается следующее по частотности (заголовку) понятие ЛО, еще не включенное в тематические узлы, и образует свой следующий тематический узел.

Приведем примеры тематических узлов, созданных в процессе обработки текста (**) из п. 3.4.2. (главное понятие тематического узла выделено сдвигом влево; указана также частота упоминания понятия в тексте):

```
1) НАРКОТИК                    3
       МОРФИН                  2
       МЕДИКАМЕНТ              1
```

2) ***БОЛЬНИЦА*** 4
ПРИЕМНОЕ ОТДЕЛЕНИЕ БОЛЬНИЦЫ 1

3) ***ПАЦИЕНТ*** 5

4) ***ВРАЧ*** 2
МЕДИЦИНСКИЙ РАБОТНИК 2

5) ***КАНАДА*** 2
АЛЬБЕРТА 1

6) ***УБИТЬ, ЛИШИТЬ ЖИЗНИ*** 1
СМЕРТЬ 2
УМЕРЕТЬ 1

7) ***ТРАВМА*** 1
НЕСЧАСТНЫЙ СЛУЧАЙ 1

8) ***МЕДСЕСТРА*** 2

В этом автоматически полученном наборе тематических узлов можно заметить следующие неточности отражения основного содержания текста.

Во-первых, тематический узел «медицинские работники» разбился на два тематических узла 4) и 7).

Возможно, правильнее иметь единый узел медицинских работников, поскольку текст делает акцент именно на вине медиков в целом:

МЕДИЦИНСКИЙ РАБОТНИК 2
ВРАЧ 2
МЕДСЕСТРА 2

Кроме того, словосочетание «несчастный случай» в тексте явно относилось не к травме, а к смерти пациента, т.е. более правильным был бы такой узел:

УБИТЬ, ЛИШИТЬ ЖИЗНИ 1
СМЕРТЬ 2
УМЕРЕТЬ 1
НЕСЧАСТНЫЙ СЛУЧАЙ 1

Но в целом, как мы видим, тематические узлы соответствуют элементам основной темы текста.

Таким образом, изложенный алгоритм формирует тематические узлы так, чтобы каждый тематический узел соответствовал отдельному элементу основной темы документа.

3.5.3.2. Мультиграфы как база для порождения тематических узлов

Как уже указывалось, построение лексических цепочек в большинстве подходов сводится, в конечном счете, к разбиению графа отношений между понятиями, упоминаемыми в тексте, на подграфы. По сути, та же процедура реализована и в процессе построения тематического представления – граф тезаурусной проекции разбивается на подграфы – совокупности тематических узлов.

Для учета факторов построения тематического представления подходит представление распределения понятий текста в виде мультиграфа, т.е. графа с двумя типами дуг между вершинами. Один тип дуг, R_{sent}, отражает отношения между понятиями в тезаурусе. Другой тип дуг, R_{text}, отражает совместную встречаемость понятий в предложениях текста. В вершинах мультиграфа указана частотность упоминания соответствующего понятия в тексте. На дугах R_{text} отмечена частота встречаемости данной пары понятий в одних и тех же предложениях текста. Дуги R_{sent} указывают частотность упоминания данной пары понятия в пределах нескольких предложений (например, 7 предложений), но не в одном предложении текста.

Таким образом, мультиграф MG тематического представления может быть определен как $MG = (V, fv, R_{text}, fr_{text}, R_{sent}, fr_{sent})$ (рис. 3.5).

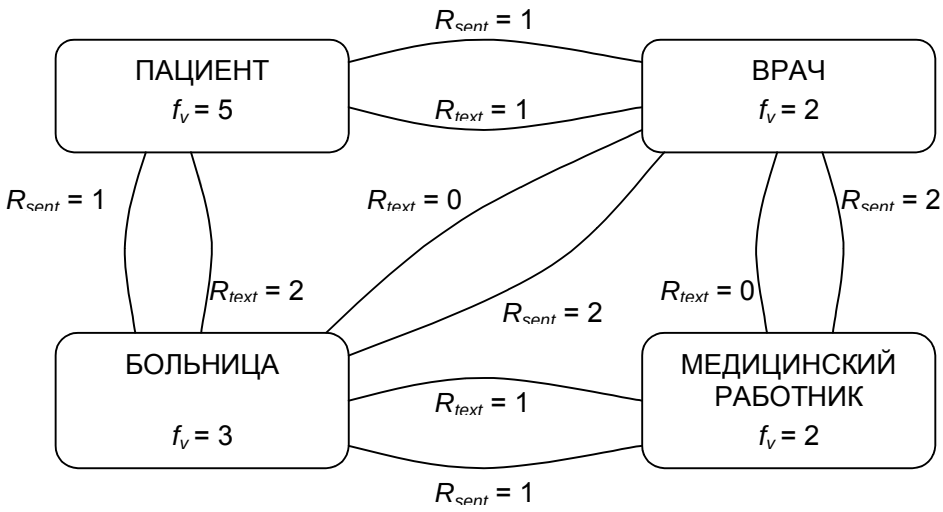


Рис.3.5. Фрагмент мультиграфа для текста (**)

3.5.3.3. Определение статуса тематического узла

На предшествующем этапе были собраны тематические узлы *Tnode},* каждый из которых включает понятия текста, связанные по отношения лингвистической онтологии ЛО с главным понятием тематического узла. С помощью тематического узла выделяются элементы основных тем и подтем текста, обсуждавшиеся в тексте.

В нашей модели мы предполагаем, что понятия основных тематических узлов постоянно встречаются рядом друг с другом (связаны по тексту) в одних и тех же предложениях текста. Понятно, что реализация проверки такого условия осложняется проблемами правильного выделения простых предложений внутри сложных предложений, построением правильной синтаксической структуры, вхождениями местоимений и использованием эллипсиса (т.е. пропусков) в тексте. Поэтому для оценки совместной встречаемости тематических узлов мы используем опять же линейный контекст понятий, называемый нами текстовые связи – r_{text}.

В результате для каждого понятия, упомянутого в тексте, получается совокупность текстовых связей, как, например, для понятия *ПАЦИЕНТ* из текста (**) (справа указана частота текстовых связей понятия *ПАЦИЕНТ* с другими понятиями текста):

ПАЦИЕНТ

НАРКОТИК	– 2
ВРАЧ	– 1
УБИТЬ, ЛИШИТЬ ЖИЗНИ	– 1
НЕСЧАСТНЫЙ СЛУЧАЙ	– 1
БОЛЬНИЦА	– 1
МЕДИЦИНСКИЙ РАБОТНИК	– 1

После того как созданы тематические узлы, текстовые связи понятий каждого тематического узла суммируются и определяются текстовые связи между тематическими узлами.

Приведем примеры текстовых связей между тематическими узлами, выделенными в тематическом представлении текста (**). Тематические узлы представлены своими главными понятиями, число справа – суммарная величина текстовых связей между понятиями тематических узлов, текстовые связи даны для тематического узла, главное понятие которого смещено в примере влево:

ПАЦИЕНТ

НАРКОТИК	– 4
БОЛЬНИЦА	– 3
ВРАЧ	– 3
УБИТЬ, ЛИШИТЬ ЖИЗНИ	– 3

…

В соответствии с рассмотрением п. 3.4.2. предполагается, что основными тематическими узлами $Tnode^M$ в первую очередь являются такие тематические узлы, которые:

- все связаны между собой текстовыми связями: $\forall\, tnode_i, tnode_j(((tnode_i \in Tnode^M) \wedge (tnode_j \in Tnode^M)) \rightarrow fr_{text}(tnode_i, tnode_j) > 0)$,

  т.е. элементы всех пар основных тематических узлов должны обсуждаться в связи с друг другом в некоторых предложениях текста;
- сумма частот текстовых связей между ними максимальна для анализируемого текста (рис. 3.6):
- $\sum_{N, i \neq j} fr_{text}(tnode^M{}_i, tnode^M{}_j) \geq \sum_{N, i \neq j} fr_{text}(tnode_i, tnode_j)$,

  где $N = \binom{m}{2}$ - число сочетаний из величины, равной количеству основных текстовых узлов $m = |Tnode^{M,}|$.

Правильно выделенные основные тематические узлы должны формировать структуру типа симплекса размерности *m-1,* и *длинами ребер,* соответствующими fr_{text} – частотностям выделенных текстовых связей, моделирующих встречаемость понятий ЛО в одних и тех же простых предложениях текста. В таком случае в силу выпуклости симплексов для любых трех тематических узлов tn_1, tn_2, tn_3 таких, что $tn_1, tn_2, tn_3 \in Tnode^M$ должны выполняться соотношения, подобные соотношению сторон в треугольнике:

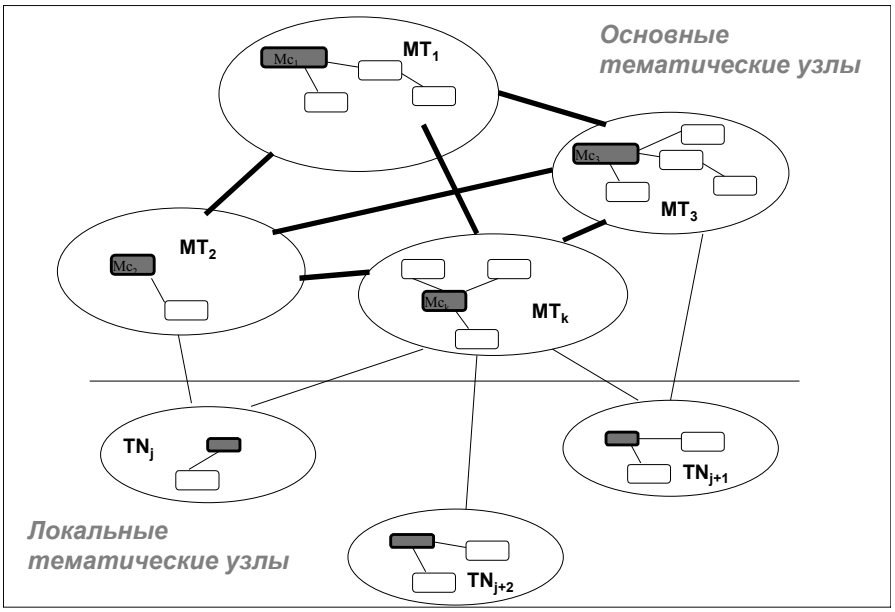


Рис. 3.6. Структура тематического представления. Линии между тематическими узлами представляют собой текстовые связи r_{text} в предложениях текста. Между основными тематическими узлами существуют наиболее частотные текстовые связи.

$$fr_{text12} < fr_{text23} + fr_{text13}$$
$$fr_{text23} < fr_{text12} + fr_{text13}$$
$$fr_{text13} < fr_{text12} + fr_{text23}$$

$$fr_{text12} > |fr_{text23} - fr_{text13}|$$
$$fr_{text23} > |fr_{text12} - fr_{text13}|$$
$$fr_{text13} > |fr_{text12} - fr_{text23}|$$

Действительно, если для некоторых tn_1, tn_2, $tn_3 \in Tnode^M$ частотности их текстовых связей образуют соотношение вида $fr_{text12} \geq fr_{text23} + fr_{text13}$, то это означает, что текстовые единицы узлов tn_1 и tn_2 образуют устойчивое словосочетание, которое должно быть собрано в единую языковую конструкцию [253], и все тематические узлы далее пересчитаны.

Если для некоторых tn_1, tn_2, $tn_3 \in Tnode^M$ частотности их текстовых связей образуют соотношение вида $fr_{text12} \leq |fr_{text23} - fr_{text13}|$, то это означает, что один из узлов tn_1, tn_2 не находится в равной степени с другими тематическими узлами в фокусе обсуждения в тексте, и, следовательно, не является основным.

Однако в реальности такие ограничения на соотношения частот текстовых связей между тематическими узлами не накладываются, поскольку пока собственно выявление тематических узлов и отслеживание их встречаемости в простых предложениях текстов отягощено большим количеством проблем, включая кореференцию, эллипсис, недостаток знаний о мире и многое другое.

В рассматриваемом примере тематического представления текста (**) основными тематическими узлами стали узлы с главными понятиями *ПАЦИЕНТ, НАРКОТИК, ВРАЧ, БОЛЬНИЦА, МЕДСЕСТРА, УБИТЬ, ЛИШИТЬ ЖИЗНИ, КАНАДА*.

Упомянутый ранее тематический узел *ТРАВМА (несчастный случай)* не прошел в список основных тематических узлов, поскольку не был связан по тексту с тематическим узлом *МЕДСЕСТРА*.

Локальные тематические узлы $Tnode^L$ представляют собой некоторые важные характеристики основных тематических узлов. Тематический узел считается локальным, если этот узел имеет текстовую связь с частотностью большей единицы с одним из основных тематических узлов. Понятия, не вошедшие в состав основных и локальных тематических узлов, объявляются "упоминавшимися" в тексте.

Таким разбиением тематических узлов *Tnode* на основные и локальные задается разбиение понятий, упомянутых в тексте, на следующие пять классов по их важности для анализируемого текста:

- главные понятия основных тематических узлов C^M (основные темы);
- другие понятия основных тематических узлов EC^M;

- главные понятия локальных тематических узлов C^L (локальные темы);
- другие понятия локальных тематических узлов EC^L;
- упоминавшиеся понятия C^O.

Таким образом, построено тематическое представление *TR* текста, в котором понятия лингвистической онтологии, упоминавшиеся в тексте, разбиты на множество тематических узлов *Tnode:* $tnode_i = (c_{0i,} \{el_1,.. el_n\})\}$. Тематические узлы подразделяются на основные $Tnode^M$, локальные $Tnode^L$ и упоминавшиеся понятия C^O. Между тематическими узлами фиксируются текстовые связи R_{text}.

3.5.3.4. Тестирование качества построения тематических узлов

В работе [109] был описан эксперимент по оценке качества тематических узлов элементов основной темы текста – макропонятий для новостных документов.

Для каждого текста человеком выбирались его основные понятия, т.е. понятия, которые наилучшим образом характеризовали основную тему анализируемого документа. Такие основные понятия выбирались, в основном, из заголовка и первого абзаца документа. Для каждого выбранного понятия автоматически строился тематический узел, состоящий из понятий данного текста.

Затем, просматривая текст, проверялось, действительно ли, включенные в тематический узел понятия, относились в данном тексте к исходному понятию. При этом было принято следующие правило: если отношение между понятиями определено в данном тексте и далее используется для организации связного текста, то это отношение не обязано быть описано в заранее созданной лингвистической онтологии, и его невключение в тематический узел не считалось ошибкой, поскольку авторы

текста не предполагали, что читатель должен знать отношения между понятиями заранее.

В нашем эксперименте все исходные макропонятия были различны и построенные узлы содержали не менее 3 понятий. На основе анализа 73 тематических узлов для 25 текстов общественно-политической тематики мы получили следующие характеристики качества отражения тематическими узлами лексической связности документов: точность – 89%, полнота – 71%.

3.5.4. Сопоставление метода построения тематического представления текстов и вероятностных тематических моделей

В последнее десятилетие в качестве моделей для статистического моделирования тематической структуры текста получили распространение так называемые статистические тематические модели (probalistic topic models), основанные на идее, что документы – это смесь тематик, где тематика – это вероятностное распределение над словами [16, 58, 59].

Тематические модели представляют собой порождающие модели (generative model). Чтобы смоделировать порождение нового документа, нужно выбрать распределение над тематиками. Затем для каждого слова в документе, случайно выбирается тематика (в соответствии с распределением) и затем случайным образом – слово этой тематики. Далее стандартные статистические методы используются, чтобы инвертировать этот процесс, и, по имеющейся коллекции текстов, восстанавливать исходный набор тематик.

Считается, что тематические модели являются развитием известного метода латентно-семантического анализа (Latent Semantic Analysis – LSA: [95]). Метод LSA основывается на трех принципах:

- семантические отношения между словами могут быть получены из встречаемости слов в документах (матрица слово-документ);

- сокращение размерности пространства важный шаг такого рода вывода,

- слова и документы могут быть представлены как точки в Евклидовом пространстве.

Статистические тематические модели базируются на первых двух предположениях, но в качестве третьего предположения принимают принцип, что семантические отношения между словами и документами могут быть описаны в виде статистических тематик. Кроме того, обе модели базируются на предположении о документе как неупорядоченной совокупности слов (bag-of words models).

Таким образом, модель определяет вероятность появления некоторого слова в тексте следующим образом:

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

где z_j – конкретная тематика, T – число тематик, $P(z)$ – вероятностное распределение над тематиками z.

В работе [16] вводится предположение об априорном распределении Дирихле для $P(z)$ – соответствующая модель была названа Скрытое распределение Дирихле (Latent Dirichlet Allocation - LDA).

На практике набор тематик для коллекции документов может моделироваться посредством применения такого варианта метода Монте-Карло, как, например, семплирование Гиббса (Gibbs Sampling). При этом в модели LDA (как и в модели LSA) должно быть сделано предположение возможном количестве тематик в коллекции (часто используется величины в интервале 50-300 тематик).

В отличие от вышеуказанных статистических моделей предложенный метод автоматического построения тематического представления текстов базируется на распределении слов и выражений не только внутри документа, но и внутри предложения, выводя необходимые соотношения из свойства глобальной связности и предикативной иерархической структуры текста. Кроме того, в представленной модели тематического представления исследуется способ использования лингвистических ресурсов в

тематическом анализе. Представленная в данной работе графовая модель тематического представления может быть впоследствии использована для создания соответствующих статистических моделей.

Заключение к главе 3.

В данной главе предложена модель представления основного содержания связного текста – тематическое представление текста. Модель опирается на предположение о том, что структура содержания текста является иерархической системой предикатов, которые соответствуют отношениям между взаимодействующими участниками описываемой в тексте ситуации. Проявление в тексте взаимодействий участников предложено описывать в виде мультиграфа с двумя типами отношений.

В результате понятия и конкретные сущности, обнаруженные в тексте, могут быть разбиты на так называемые тематические узлы, каждый из которых соответствует отдельной стороне (группе участников), взаимодействующих в ситуации.

Такая модель позволяет уйти от распространенной в современных информационных системах модели содержания текста как совокупности независимых друг от друга слов («мешок слов») и предложить способ определения значимости (веса) понятия в тексте с учетом близких по смыслу понятий – элементов того же тематического узла.

Глава 4. Автоматическая обработка текстов на основе лингвистической онтологии и приложения информационного поиска

4.1. Этапы обработки текстов на основе ЛО

Автоматическая обработка текстов на основе лингвистической онтологии ЛО состоит из следующих этапов:

1) производится графематический и морфологический анализ текста;

2) текст сопоставляется с единицами ЛО. Из множества найденных в конкретном месте текста единиц ЛО выбирается единица, имеющая максимальную длину. Если один и тот же фрагмент текста соответствует разным единицам ЛО, то фиксируется многозначность термина. В результате сопоставления с ЛО текст отражается в последовательность понятий ЛО. Все синонимы (варианты) одного и того же понятия отображаются в соответствующий номер понятия и далее не различаются (рис. 4.1);

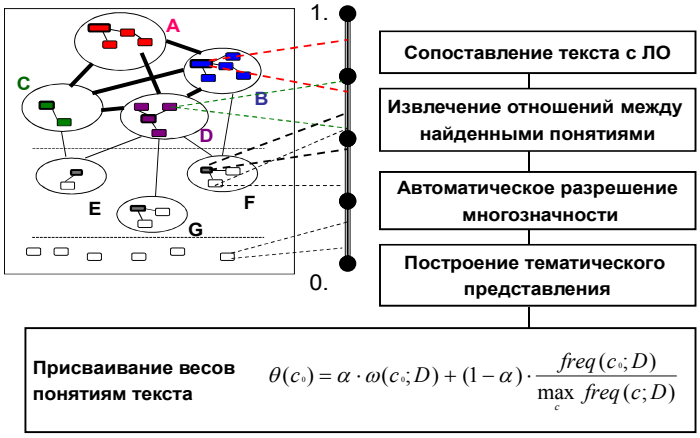


Рис. 4.1. Схема построения тематического представления и концептуального индекса документа

3) строится т.н. проекция онтологии на текст, что представляет собой создание фрагмента онтологии, соответствующего данному тексту. В данный фрагмент онтологии переносятся понятия онтологии, упомянутые в тексте, и отношения между ними как прямо указанные в онотологии, так и выводимые

по свойствам отношений. Для большинства текстов, такой фрагмент онтологии представляет собой сложную сеть отношений, которая может распадаться на несколько несвязанных подфрагментов, а может содержать достаточно много различных связанных между собой понятий. На рис. 4.3 представлен фрагмент онтологии для текста (рис. 4.2);

О порядке оказания ***безвозмездной финансовой помощи*** на ***строительство (покупку) жилья*** и ***выплаты денежной компенсации*** за ***наем (поднаем) жилых помещений военнослужащим*** и ***гражданам, уволенным*** с ***военной службы***

Во исполнение ***Закона Российской Федерации*** "О ***статусе военнослужащих***" и в целях обеспечения ***прав*** на ***жилище военнослужащих*** и ***граждан, уволенных*** с ***военной службы, Правительство Российской Федерации*** п о с т а н о в л я е т :

1. ***Утвердит***ь прилагаемое Положение о порядке оказания ***безвозмездной финансовой помощи*** на ***строительство (покупку) жилья*** и ***выплаты денежной компенсации*** за ***наем (поднаем) жилых помещений военнослужащим*** и ***гражданам, уволенным с военной службы.***

2. ***Министерству обороны Российской Федерации*** и иным ***федеральным органам исполнительной власти,*** в которых предусмотрена ***военная служба***:

в месячный срок разработать и утвердить формы и перечень ***документов,*** необходимых для принятия решения об оказании ***военнослужащим безвозмездной финансовой помощи*** на ***строительство (покупку) жилья*** и о ***выплате денежной компенсации*** за ***наем (поднаем) жилых помещений;***

расходы, связанные с оказанием ***военнослужащим безвозмездной финансовой помощи*** и ***выплатой денежной компенсации*** за ***наем (поднаем) жилых помещений,*** производить за счет и в пределах ***средств,*** выделяемых из ***федерального бюджета*** по ***сметам*** этих ***федеральных органов исполнительной власти.***

3. ***Органам исполнительной власти субъектов Российской Федерации***:

оказывать ***безвозмездную финансовую помощь*** в ***избранном постоянном месте жительства гражданам, уволенным***

Рис. 4.2. Пример текста (Постановление Правительства РФ от 26 июня 1995 г. № 604) с выделенными терминами Общественно-политического тезауруса.

4) для многозначных текстовых единиц производится автоматическое разрешение многозначности;

5) Далее на базе выявленных в тексте понятий онтологии строится тематическое представление текста, которое позволяет сгруппировать понятия текста в группы близких по смыслу – тематические узлы и выявить значимость каждого тематического узла.

6) Полученное тематическое представление трансформируется в концептуальный индекс текста, в котором каждое понятие присвоен вес (рис. 4.1).

Построенный концептуальный индекс используется в разнообразных приложениях таких, как концептуальное индексирование документов, автоматическая рубрикация, автоматическое аннотирование [3, 287, 288].

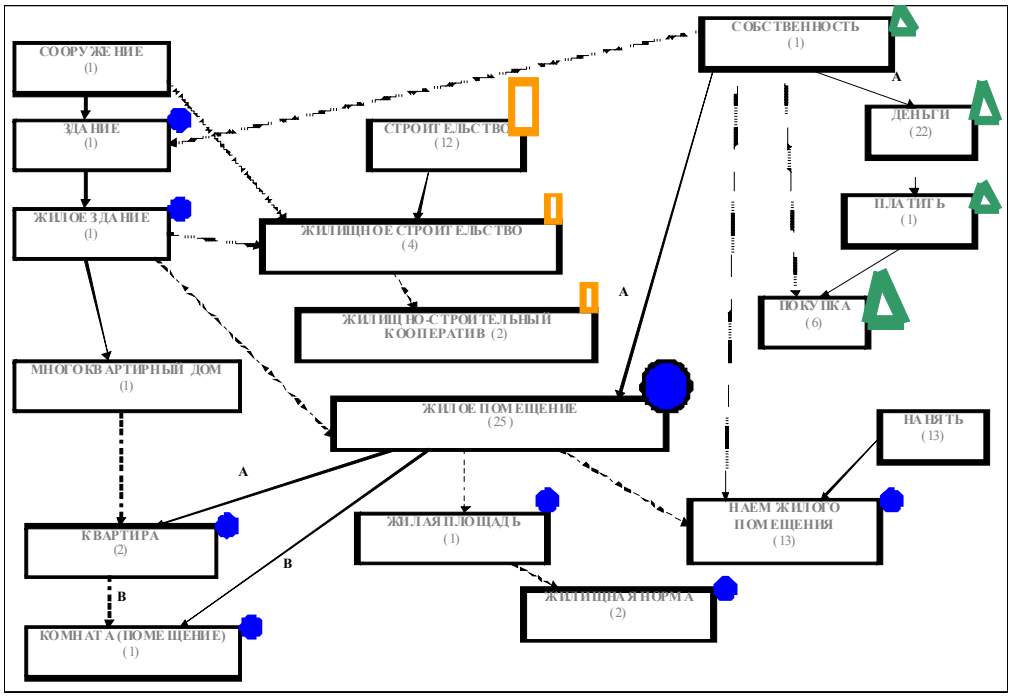


Рис. 4.3. Фрагмент понятийной сети для
текста Постановления Правительства РФ от 26 июня 1995 г. № 604

4.2. Автоматическое разрешение многозначности

Существенной проблемой при автоматической обработке текстов является многозначность текстовых единиц, представленных в онтологии.

Поэтому важно развитие качественного алгоритма автоматического разрешения многозначности [272, 291, 293].

Согласно модели лингвистической онтологии, представленной в главе 2, имеется два способа представления значений многозначных выражений t_m в ЛО: $t_m \in M,\ M = M_m \cup M_a$.

Первым способом представления многозначности является задание одного и того же текстового входа разным понятиям ЛО (М-многозначность): $t_m \in M_m$. Так, например, текстовый вход *пилот* в тезаурусе РуТез сопоставлен двум разным понятиям: понятию *ЛЕТЧИК* и понятию *АВТОГОНЩИК*.

Второй способ представления многозначности используется в тех случаях, когда слово представлено в ЛО 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): $t_m \in M_a$.

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).

Основным подходом разрешения лексической многозначности на основе лингвистической онтологии является сопоставление контекста понятий, соответствующих значениям многозначного слова, в анализируемом тексте и их окрестности в ЛО.

Таким образом, в задаче разрешения многозначности имеются следующие исходные данные:

- многозначному текстовому входу ЛО $t \in T$ соответствует совокупность понятий ЛО $C^{t=}\{c_1^t .. c_n^t\}$,

- для каждого понятия онтологии определена окрестность близких понятий $O(c_i)$ (см. п.2.3), совокупность всех окрестностей для множества C^t обозначим как C^{tokr},

- текстовый вход t упоминается в контексте, в котором выявлены некоторое множество понятий онтологии $C^{context=}\{c_1^{context},.. c_k^{context}\}$.

Для разрешения многозначности многозначного текстового входа t необходимо построить функцию выбора между значениями F_{disamb}:

$$F_{disamb}: C^t \times C^{context} \times C^{tokr} \rightarrow c_i^t$$

которая на основе списка понятий C^t, соответствующих многозначному слову t, контекста употребления слова t в тексте $C^{context}$ и окрестностей понятий C^{tokr} в ЛО, выбирает $c^t \in C^t$ так, что это максимально соответствует выбору экспертов. Фактически функция выбора значения F_{disamb} является функцией определения максимального сходства между окрестностью понятия в лингвистической онтологии и контекстом в тексте:

$$F_{disamb} = argmax_i\ (Fsim\ (C^{context}, C_i^{tokr}))$$

При этом контекст употребления слова может быть как локальным: несколько слов влево и вправо, так и глобальным - весь текст:

$$C^{context} = C^{loc} \cup C^{glob}$$

4.2.1. Метод глобального подтверждения разрешения лексической многозначности

Метод глобального подтверждения заключается в том, что все понятия ЛО, вхождения которых обнаружены в тексте, могут оказывать влияние на выбор значения многозначного языкового выражения [273]. Рассмотрение глобального контекста учитывает такое свойство связного текста как лексическую связность текста, т.е. повторяемость одних и тех же лексических единиц и совокупностей семантически близких лексических единиц в связном тексте.

Для каждого варианта многозначного выражения $t_m \in M$ собираются те понятия текста, которые "поддерживают" этот вариант. Пусть t_m в одном из значений относится к понятию c_i, т.е. s (t_m, c_i). "Поддержка" в тексте T этого значения может осуществляться двумя способами:

- в тексте встречается однозначный текстовый вход понятия c_i, т.е. $\exists\ t_0$: $(t_0 \in T \wedge t_0 \notin M \wedge s\ (t_m, c_i) \wedge s\ (t_0, c_i))$, например, упоминание в тексте словосочетания *расследование преступлений* поддерживает именно это значение у многозначного слова *следствие;*
- в тексте упоминается понятие c_j из окрестности $O\ (c_i)$ (см. п. 2.5.1), т.е. $\exists\ t_0$: $(t_0 \in T \wedge s\ (t_m, c_i) \wedge s\ (t_0, c_j) \wedge c_j \in O\ (c_i))$, например, упоминается понятие *ОБЩЕСТВЕННАЯ ДЕЯТЕЛЬНОСТЬ* из окрестности понятия *ПОЛИТИЧЕСКАЯ ПАРТИЯ*, к которому относится неоднозначный термин *партия.*

Далее собственно и производится выбор варианта понятия для многозначного термина. Поскольку многозначность в ЛО может быть задана двумя способами: с помощью пометы и с помощью отнесения текстового выражения к разным понятиям ЛО, то процедура автоматического выбора значения в этих случаях несколько различается:

- неоднозначность задана с помощью пометы, т.е. $t_m \in M_a$. Если текст "поддерживает" описанное в ЛО значение неоднозначного термина, то соответствующее понятие включается в понятийный индекс как однозначный. В противном случае, неоднозначный термин исключается из концептуального индекса текста;
- неоднозначность проявляется в соответствии одного текстового выражения нескольким понятиям, т.е. $t_m \in M_m$. Сначала проверяется, какие из вариантов термина поддерживаются понятиями всего текста, и оставляются только "поддержанные" варианты. Если ни один из вариантов не поддерживается, то все они удаляются из концептуального индекса.

После удаления "неподдержанных" вариантов может остаться только один вариант, и, таким образом, неоднозначность разрешена.

Если же поддержано более одного варианта, то производится выбор значения именно для конкретного вхождения неоднозначного термина: выбирается тот вариант, для которого "поддерживающее" понятие находится ближе всего по тексту. Расстояние измеряется в количестве выявленных понятий между текущим вхождением неоднозначного термина и "поддерживающим" понятием.

Далее этот метод разрешения многозначности мы будем называть Glob.

Данный алгоритм очень прост, и в нем есть некоторые проблемы.

Во-первых, в этом методе для учета концептуальной близости используются только пути P_{up} и P_{down} (см. п. 2.5.1), состоящие из иерархических отношений одной направленности, т.е. 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, нет ограничений на длину пути между понятиями, что может привести, например, к тому, что многозначность очень конкретного понятия могла быть разрешена на основе нахождения в тексте очень абстрактного понятия.

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

В-четвертых, приоритет отдается глобальному контексту, т.е. сначала проверяется, если ли подтверждение для того или иного значения по всему тексту. Если несколько значений имеют подтверждение в глобальном контексте, то проверяется локальный контекст: выбирается то значение, подтверждение для которого находится ближе всего к исследуемому многозначному вхождению.

Поэтому был предложен другой алгоритм разрешения многозначности, который более аккуратно учитывает разные характеристики путей между понятиями ЛО.

4.2.2. Метод взвешивания подтверждения от локального и глобального контекстов

Основой для нового разработанного алгоритма разрешения многозначности является оценка семантической близости *Fsim* между возможными значениями, с одной стороны, и окружающим текстовым контекстом, с другой стороны [291, 293]. При этом рассматривается как локальный контекст C^{Loc}, который задается в виде некоторого окна – линейной окрестности многозначного вхождения слова, так и глобальный контекст C^{Glob}, в который входят все понятия ЛО, упомянутые в тексте.

Функция *Fsim* имеет очень сложную природу – в данной работе она моделируется как линейная функция от большой совокупности факторов; вклад каждого фактора моделируется как коэффициент, подбираемый на основе метода покоординатного спуска.

4.2.2.1. Учет локального и глобального контекста

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

Кроме того, исследовалось задание локального контекста как «динамического» окна N+N, т.е. 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 [52] разные типы связи имеют разную сферу действия и разный вес в зависимости от такого рода расстояния, измеряемого в абзацах и предложениях.

В процессе экспериментов была выбрана следующая специфика учета глобального контекста.

В качестве элементов глобального контекста учитываются только однозначные вхождения единиц ЛО: $t \notin M$. При этом никаких ограничений на расстояние между вхождением многозначного слова и семантически близкими словами не накладывается. 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 необходимо сбалансировать свидетельства в пользу того или иного значения, получаемые от локального и глобального контекста. 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, была протестирована возможность применения коэффициента, уменьшающего вес подтверждения от глобального контекста при увеличении длины текста (точнее при увеличении максимальной частотности лексической единицы в тексте).

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

Семантическая близость между двумя понятиями c_1 и c_2 оценивается на основе рассмотрения пути отношений, который существует между этими единицами ЛО.

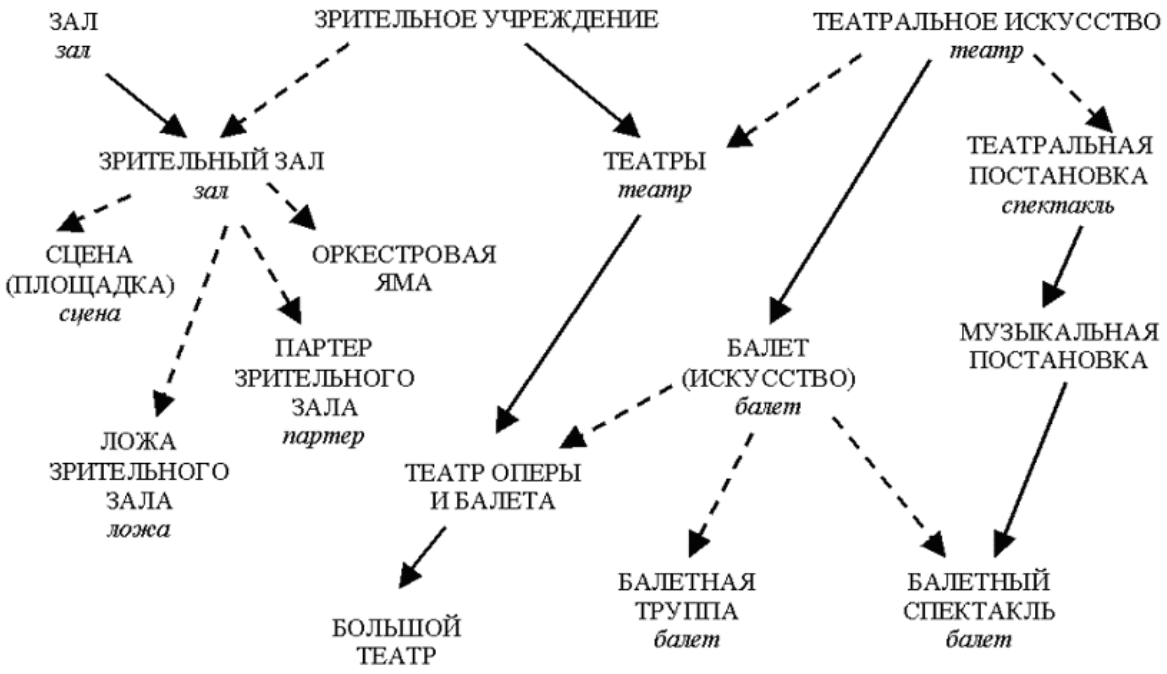


Рис. 4.4. Фрагмент сети понятий текста с многозначными текстовыми входами

Между понятиями в ЛО могут существовать пути разной конфигурации, ЛО является связной и всегда существует путь отношений от одного произвольного понятия ЛО до другого понятия ЛО. Однако подобно подходу [76] мы ограничиваем конфигурации путей между понятиями c_1 и c_2, которые рассматриваются при оценке семантической близости понятий, а именно, либо путь должен состоять из совокупности иерархических отношений, направленных в одну сторону (пути P_{up} и P_{down}), например, последовательность отношений от вида к роду (иерархический путь), либо

такой путь должен включать ровно один перегиб, т.е. 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 (пути P_{updown} и P_{downup} – см. п. 2.3).

На рис. 4.4 примером иерархического пути P_{up} является путь

БОЛЬШОЙ ТЕАТР

– (выше) – ТЕАТР ОПЕРЫ И БАЛЕТА

– (целое) – БАЛЕТ (ИСКУССТВО),

Примером пути с перегибом сверху P_{updown} является путь:

ОРКЕСТРОВАЯ ЯМА

– (целое) – ЗРИТЕЛЬНЫЙ ЗАЛ

– (часть) – ПАРТЕР ЗРИТЕЛЬНОГО ЗАЛА,

Примером пути с перегибом снизу P_{downup} является путь:

ТЕАТРАЛЬНАЯ ПОСТАНОВКА

– (ниже) – БАЛЕТНЫЙ СПЕКТАКЛЬ

– (выше) – МУЗЫКАЛЬНАЯ ПОСТАНОВКА.

Ограничение просмотра путей между понятиями именно такими типами связано с тем, что любой иерархический путь P_{up} или P_{down} мужду понятиями правилами транзитивности и наследования (см. п. 2.3) может быть сведен к пути длиной в одно отношение, а пути с перегибами P_{updown} и P_{downup} – к пути длиной в два отношения. Таким образом, доказывается потенциальная близость понятий, соединенных путями P_{up}, P_{down}, P_{updown} и P_{downup}.

4.2.2.3. 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:

$$F_{Sim}(c_1, c_2) = максимальный\_балл - длина\_пути - цена\_многозначности - цена\_перегиба - цена\ глобальности. \quad (4.1)$$

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, большую нуля, в случае оценки глобального контекста, и величину, равную нулю, при анализе локального контекста.

Таким образом, алгоритм имеет следующий набор параметров:

- максимальная длина дерева, т.е. насколько далеко в одном и то же направлении иерархических отношений от исходного понятия можно искать подтверждающие значение понятия – длина дерева может быть различной для локального и глобального контекстов,
- строение (статическое или динамическое) 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 контекста,
- пороги для видов многозначности: А-многозначности и М-многозначности.

Параметры алгоритма подбирались на основе размеченной значениями коллекции текстов *методом покоординатного спуска*.

4.2.2.4. Этапы алгоритма

Поступающий текст проходит через процедуру графематического и морфологического анализа. Далее на основе цепочек лемм, полученных в результате морфологического анализа, происходит сопоставление с ЛО. Для каждого сопоставившегося текстового входа ЛО отмечается его статус: однозначное сопоставление, сопоставление с пометкой многозначности

(А-многозначность), сопоставилось несколько текстовых входов ЛО (М-многозначность). Отметим, что если один из сопоставленных текстовых входов ЛО, полностью включается в другой, более длинный текстовый вход, то эта ситуация многозначной не считается, сопоставленным считается более длинный текстовый вход.

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 (4.1). Все набранные баллы понятий-значений многозначных единиц суммируются и запоминаются.

Далее происходит анализ локального контекста. Для каждого вхождения многозначной текстового входа ЛО просматривается заданная текстовая окрестность, выбираются упоминаемые понятия, связанные с понятиями данной многозначной единицы путями разрешенной конфигурации, и подсчитываются баллы по формуле (4.1). Баллы, полученные при глобальном анализе и локальном анализе, 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 другой 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. 4.4*)*. В случае если такой иерархической связи не имеется, то в настоящее время не выбирается

ни одно из понятий – многозначность остается неразрешенной. Если бы на основе разметки корпуса было бы известно наиболее частотное значение, то можно было бы в таких случаях выбирать именно это частотное значение.

Далее на этот алгоритм разрешения многозначности мы будем ссылаться LocGlob. Всего в алгоритме LocGlob используется 47 параметров (табл. 4.1), включая различные конфигурации путей между понятиями в онтологии. Параметры алгоритма подбирались на основе размеченной лексическими значениями коллекции текстов *методом покоординатного спуска.*

Табл.4.1. Список параметров метода LocGlob разрешения многозначности

Величина локального окна
Динамическое увеличение локального окна
Глубина деревьев на локальном уровне
Глубина деревьев на глобальном уровне
Цена пометок на отношениях (локальный уровень)
Цена верхнего уровня (локальный уровень)
Цена многозначности (локальный уровень)
Делитель/вычитаемое (локальный уровень)
Порог А-многозначности (локальный уровень)
Порог М-многозначности (локальный уровень)
Цена пометок на отношениях (глобальный уровень)
Цена верхнего уровня (глобальный уровень)
Цена глобальности (глобальный уровень)
Умножение на частотность (глобальный уровень)
Порог А-многозначности (глобальный уровень)
Порог М-многозначности (глобальный уровень)
Цена перегиба по отношениям – 28 параметров в зависимости от типов отношений и уровня (глобальный, локальный

4.2.3. Организация тестирования алгоритмов разрешения многозначности

Для определения качества разрешения лексической многозначности необходимо было выполнить эталонную разметку найденных в тексте входов ЛО по значениям. Для каждого документа экспертами-лингвистами были созданы эталонные файлы, с правильной разметкой значений [293].

После получения эталонных файлов они автоматически сопоставляются 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 ЛО, отмеченных как многозначные.

Тестировалась отдельно точность разрешения многозначности по Общественно-политическому тезаурусу, т.е. определялось качество разрешения многозначности тематической лексики и терминологии, и по тезаурусу РуТез, т.е. тестировалось качество разрешения многозначности для всех знаменательных слов текста. Последняя задача соответствует задаче

тестирования «все слова текста», проводимой в рамках конференции Senseval.

4.2.3.1.Тестирование алгоритмов разрешения многозначности на основе Общественно-политического тезауруса

Тестирование алгоритмов 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 (данная коллекция распространяется в рамках экспериментов семинара РОМИП).

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 алгоритмов разрешения многозначности по каждому из источников показаны в Табл. 4.2, где N_{doc} – число документов, N_{amb} – число вхождений неоднозначных терминов, $P_{locglob}$ – точность по алгоритму LocGlob, P_{glob}– точность по алгоритму Glob.

Источник	N_{doc}	N_{amb}	$P_{glob+loc}$, %	P_{glob}, %
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

Табл 4.2. Точность разрешения лексической многозначности по источникам публикаций

Совокупная точность работы системы по более гибкому алгоритму LocGlob в процессе тестирования составила 73.37% и выросла на 6.7% относительно точности разрешения многозначности, полученной по алгоритму Glob.

Как и предполагалось, наибольший рост точности алгоритма, более гибко учитывающего конфигурации путей отношений тезауруса, а также локальный и глобальный контекст, удалось получить на относительно коротких текстах новостных сообщений. Рост точности разрешения многозначности на этих типах текстов составил более 10%.

Для получения лучших результатов тестировались разные наборы параметров алгоритма LocGlob.

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, т.е. 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*.

Также исследовался 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.

4.2.3.2. Тестирование алгоритма разрешения многозначности по Тезаурусу РуТез

Для тестирования алгоритма разрешения многозначности по всему Тезаурусу РуТез, что соответствует задаче «все слова текста» конференции Senseval, было взято по 2 статьи из газет «Известия», «Комсомольская правда», «Независимая газета», «Ведомости». Количество многозначных единиц – 1120. Меньший объем коллекции объясняется значительно большими трудозатратами по подготовке эталонной разметки. Для алгоритма LocGlob была получена точность разрешения многозначности – 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 (чем на основе Общественно-политического тезауруса) – используется динамическое окно 4+4.

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

Данный результат был получен 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.

Заключение к разделу 4.2

Таким образом, реализованные алгоритмы автоматического разрешения многозначности показали максимальную среднюю точность разрешения многозначности 73.37% для тематической лексики и терминологии Общественно-политического тезауруса, и 57.14% для всех знаменательных слов текста, т.е. по тезаурусу РуТез в целом. Качество разрешения многозначности для задачи «все слова текста» значительно превышает показатели, достигнутые для алгоритмов, работающих на основе WordNet в тех же условиях, т.е. без учета информации из размеченного корпуса, и, в частности информации о самом частотном значении. Это, на наш взгляд, в значительной мере связано с более богатой отношениями структурой ЛО.

Существенной особенностью реализованного подхода к автоматическому разрешению лексической многозначности является учет совокупности различных контекстных факторов, и для нахождения их оптимальной комбинации был использован численный метод оптимизации.

4.3. Информационный поиск на базе ЛО

4.3.1. Концептуальный индекс, веса понятий и отношений

Тематическое представление текста дает возможность построить концептуальный индекс документа, в котором учитывается не только частотность отдельного понятия в документе, но и статус понятия в тематической структуре документа. В результате построения тематического представления текста все понятия ЛО, упомянутые в тексте, разделяются на пять базовых классов значимости для текста, каждый из которых имеет свой вес. Задание весов этих классов может осуществляться параметрически. В большинстве случаев, веса классов значимости понятий задаются следующим образом [250]:

- центры основных тематических узлов – 0.95;
- другие понятия основных тематических узлов – 0.85;
- центры локальных тематических узлов – 0.70;
- другие понятия локальных тематических узлов – 0.65;
- упоминавшиеся понятия, не вошедшие в предыдущие классы – 0.20.

Базовый вес понятия получен в качестве интегрального анализа распределения в тексте совокупностей близких по смыслу терминов. Чтобы снизить фактор ошибки вычисления базовых весов, а также сделать веса понятий более дробными, для формирования окончательного веса понятий учитывается также относительная частотность понятий в тексте. Окончательный вес понятия в тексте $\mu(c, \mathrm{D})$ рассчитывается по следующей формуле:

$$\mu(c, \mathrm{D}) = \lambda \cdot \nu^*(c, \mathrm{D}) + (1-\lambda) \cdot freq(c, \mathrm{D}) \cdot [freq^*(\mathrm{D})]^{-1} \qquad (4.2)$$

где $\nu^*(c, \mathrm{D}) = \max_{Th(c,\ \mathrm{D})} \nu(\mathrm{c}, \mathrm{D})$ – максимум базовых весов понятия c в тематических узлах; $freq(c, \mathrm{D})$ – частота понятия c в документе D, $freq^*(\mathrm{D}) = \max_{d \in \mathrm{D}} freq(d, \mathrm{D})$ – максимальная частота среди понятий

документа D, λ – подбираемый параметр формулы (обычно устанавливается равным 0.7). Таким образом, при загрузке текстов в поисковую систему создается концептуальный индекс текста по ЛО, строится тематическое представление текста, каждому понятию присваивается вес по формуле (4.2).

При расширении запроса по ЛО необходимо организовать выдачу и таких текстов, в которых нет исходных понятий запроса, но имеются понятия нижестоящие по иерархии – для этого используются понятия из нижней окрестности понятия (или дерево расширения вниз) O^- (см. п. 2.3).

Каждое понятие в дереве расширения имеет свой вес, который зависит от суммарного отношения данного понятия к исходному понятию и не зависит от длины пути до понятия-вершины дерева. Эти величины используются как коэффициенты, на которые домножается вес, присвоенный данному понятию при анализе конкретного документа.

Документ может содержать несколько различных понятий из дерева расширения. Для вычисления веса такого документа веса всех понятий из дерева расширения суммируются так, чтобы придать больший вес документам, которые содержат несколько понятий из дерева расширения:

$$\mu^+(c,D) = 0.7 \cdot \max_{c_i \in O^-(c)} \{\mu(c_i)\cdot\} + 0.3 \cdot \max\left\{\mu(c), \frac{R(c_i,D)}{1+R(c_i,D)}\right\}, \quad (4.3)$$

где $R(c,D) = \sum_{c_i \in O^-(c)} V(c_i)$ – сумма весов понятий $c_i \in O^-(c)$, упомянутых в документе D.

4.3.2. Тестирование эффективности информационного поиска на основе ЛО

Для проверки работы модели лингвистической онтологии при расширении запросов в процессе информационного поиска был выполнен эксперимент по тестированию качества поиска при использовании Общественно-политического тезауруса [250]. Исследовались запросы, которые соответствуют текстовым входам Общественно-политического тезауруса, и, таким образом, изучалась работа модели описания отношений в ЛО и ее реализация в Общественно-политическом тезаурусе. В качестве запросов случайным образом были выбраны рубрики из Классификатора правовых актов [311]. Поиск осуществлялся на коллекции нормативных актов УИС Россия [226].

Для тестирования эффективности информационного поиска мы выполнили набор запросов в УИС РОССИЯ. Каждый запрос был сформулирован дважды: один раз как запрос на поиск по словам, второй раз – как запрос на поиск по понятиям тезауруса с полным расширением по дереву. Поиск по словам осуществляется с использованием векторной модели в формулировке системы Inquery [26]. Вес слова tf.idf векторной модели вычисляется по следующим формулам:

$$tf_D(l) = \frac{\text{freq}_D(l)}{\text{freq}_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{\text{avg\_dl}}}$$

$$idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)}$$

где $freq_D(l)$ – частота леммы l в документе D, dl_D – мера длины документа, в нашем случае количество лемм в документе, avg_dl – средняя длина документа, $df(l)$ – количество документов в коллекции, где встречалась лемма l, $\beta = 0.4$, $|c|$ – количество документов в коллекции.

При выполнении подавляющего количества запросов количество документов, найденных с использованием деревьев Общественно-политического тезауруса значительно превышало количество документов, найденных по словам. Таким образом, полнота поиска с использованием деревьев тезауруса значительно возросла. Однако, как известно, увеличение полноты поиска часто сопровождается снижением точности поиска, т.е. релевантными считается большее количество нерелевантных документов.

Чтобы сопоставить точность поиска по тезаурусу и по словам, мы использовали методику оценки средней точности по трем заданным значениям полноты, описанную в [209]. Точность выполнения запроса вычисляется при следующих трех значениях полноты: 0.2, 0.5, 0.8.

Чтобы оценить эффективность поиска, необходимо сначала определить множество релевантных документов, а затем проверить релевантность значительного количества полученных по запросу документов. Для снижения трудозатрат, необходимых на проведение оценок, сокращался временной интервал до тех пор, пока не было получено 30-40 документов в поисковой выдаче. Эффективность поиска на таком количестве документов уже достаточно просто проверить.

Всего было выполнено тестирование 19 запросов – рубрик Классификатора нормативных актов [311] (см. рис. 4.4, 4.5).

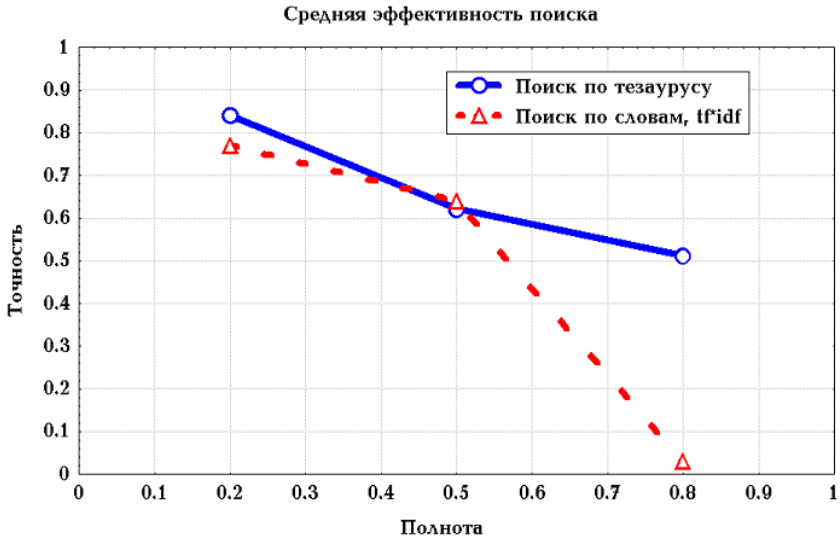
Точность по терминам:

- Точность по терминам в точке 0.2: – 15.41 – 0.81
- Точность по терминам в точке 0.5: – 11.04 – 0.58
- Точность по терминам в точке 0.8: – 8.80 – 0.46
- Средняя точность: = 0.62

Точность по словам:

- Точность по терминам в точке 0.2: – 14.76 – 0.77
- Точность по терминам в точке 0.5: – 9.77 – 0.52
- Точность по терминам в точке 0.8: – 0.36 – 0.02

- Средняя точность: = 0.44



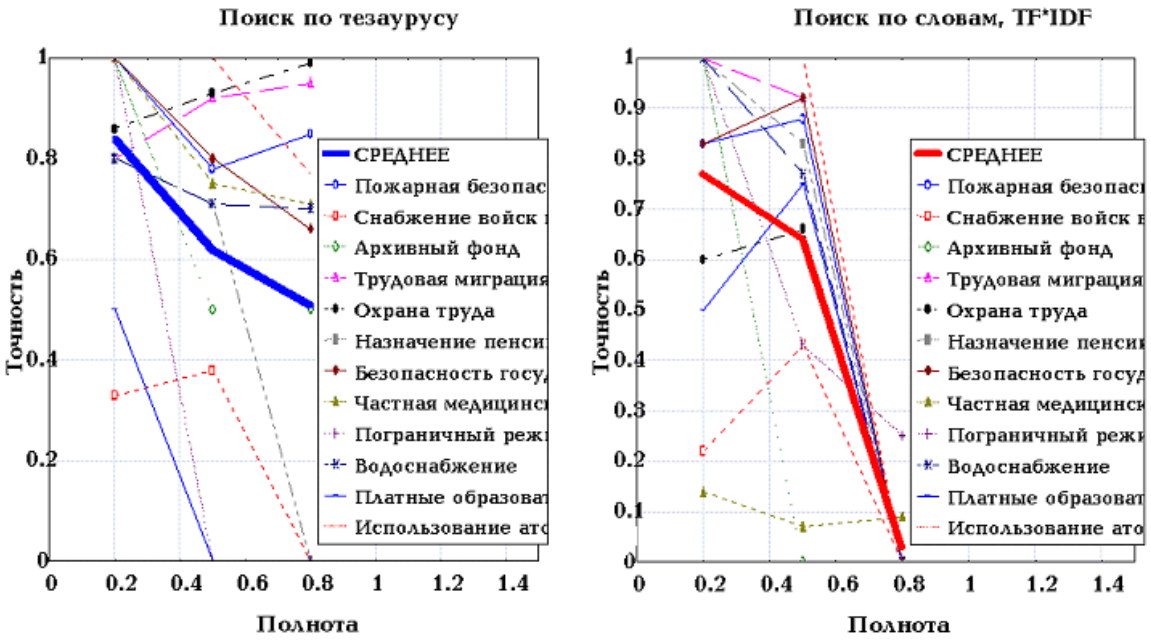


Рис. 4.4, 4.5 Графики средней точности поиска
по Общественно-политическому тезаурусу и по пословной векторной модели
в базе нормативных актов УИС РОССИЯ

Отметим, что в условиях эксперимента запросы были небольшой длины и при этом имели достаточно хорошее пересечение с терминами Общественно-политического тезауруса. На практике частой ситуацией

является наличие в запросе большого количества слов, не входящих в Общественно-политический тезаурус, имеющих другое значение, чем описано в Общественно-политическом тезаурусе и др.

Данный эксперимент подтверждает, что при совпадении запроса с термином тезауруса расширение поиска по тезаурусу приводит к значительному увеличению эффективности информационного поиска. Кроме того, этот эксперимент подтверждает, что наши усилия описывать наиболее надежные, применимые в разных контекстах, отношения в тезаурусе также дали свои результаты.

4.3.3. Лингвистическая онтология и векторная модель в задаче поиска по коллекции нормативно-правовых актов РОМИП

В реальных условиях задания запросов пользователем запросы по отношению к лингвистической онтологии могут быть весьма разнообразны:

- запрос может быть очень коротким (например, содержать отдельное многозначное слово, значение которого без диалога с пользователем выяснить невозможно),
- запрос может содержать некоторую совокупность слов, в которой не найдены единицы ЛО,
- запрос может быть достаточно длинным, и одна часть запроса может ограничивать контекст расширения для другой части запроса и др.

Для учета разных ситуаций была предложена смешанная модель, основанная на совокупности факторов, включая веса слов по пословной векторной модели, веса понятий лингвистической онтологии, нахождение сущностей из запроса в ограниченном числе предложений документа. Модель тестировалась на семинаре РОМИП-2008 в коллекции нормативно-правовых документов [228], в качестве лингвистической онтологии использовался Общественно-политический тезаурус.

Основной направленностью разработки модели была обработка длинных информационных запросов, т.е. запросов, которые имеют длину более 3 слов, и выражают некоторую информационную потребность. Информационные запросы условно противопоставляются навигационным запросам, суть последних в нормативно-правовой коллекции заключается в получении документа путем задания его формальных реквизитов: типа документа, номера документа, даты выхода, заголовка.

Для поиска документов по запросам в нормативно-правовой коллекции использовалась двухшаговая процедура.

На первом этапе исполнялась комбинированная векторная модель, построенная на двух индексах – индексе лемм и индексе понятий Общественно-политического тезауруса.

Понятия тезауруса дают возможность дополнительно учесть три дополнительных фактора:

- синонимию терминов,
- лексическую многозначность – производится предварительный выбор наиболее подходящего по контексту значения слов и выражений,
- близкое расположение в тексте компонентов многословных терминов и выражений.

Поэтому результаты работы двух видов векторных моделей могут достаточно серьезно различаться.

Результаты работы векторных моделей замешиваются с помощью параметра α_1, т.е. каждый документ получает вес по следующей формуле:

$$W_d = \alpha_1 W_{word} + (1 - \alpha_1) W_{conc,} \qquad (4.6)$$

где W_{word} – вес документа по пословной векторной модели, W_{conc} – вес документа по векторной модели, выполненной на основе концептов тезауруса.

Из документов, найденных по смешанной векторной модели, отбирается 100 документов.

На втором этапе обработки запроса найденные 100 документов переупорядочиваются по следующему принципу. Максимальное число элементов запроса (слов и терминов) должно быть найдено не разбросанными по всему тексту, а сосредоточены в двух парах соседних предложений. Коэффициент α_2 оценивает относительную весовую значимость лемм и понятий тезауруса в предложениях.

Получение нового веса документа можно представить как двухпроходный процесс. Сначала подсчитываются веса отдельных предложений, которые получаются суммированием весов лемм и концептов из запроса, найденных в предложении:

$$W_s = \alpha_2 \Sigma w_{wordi} + (1 - \alpha_2) \Sigma w_{concj} \qquad (4.7)$$

где w_{wordi}, w_{concj} – веса слов и концептов предложения.

На втором проходе вычисляется «усиленный» вес каждого предложения: если не все элементы запроса найдены в текущем предложении, то проверяется, нет ли недостающих элементов в соседнем предложении или в еще одной паре предложений документа. Веса дополнительных элементов найденных в других предложениях домножаются на параметрические коэффициенты α_4 (для присоединения элементов из соседнего предложения) и α_5 (для присоединения элементов из другой пары рядом лежащих предложения).

Таким образом, формула «усиленного» веса предложения имеет следующий вид:

$$W_{s1+} = W_1 + \alpha_4 W_{2-} + \alpha_5 [W_{3-} + \alpha_4 W_{4-}], \qquad (4.8)$$

где W_1 – вес «главного» предложения, W_{2-} – вес следующего предложения, W_{3-}, W_{4-} – веса еще одной пары смежных предложений. Причем для каждого следующего предложения учитываются только те слова и концепты,

ассоциируемые с запросом, которые еще не были учтены для предыдущих предложений.

Наконец, **на третьем этапе** исходный вес документа**,** полученный на первом этапе, замешивается с весом документа по предложениям, полученный на втором этапе.

Параметры модели оптимизировались на материалах дорожки нормативно-правового поиска romip-legal-2005 *методом координатного спуска*. Оптимизировалось максимальное число релевантных документов в первых пяти документах выдачи, т.е. показатель Precision(5).

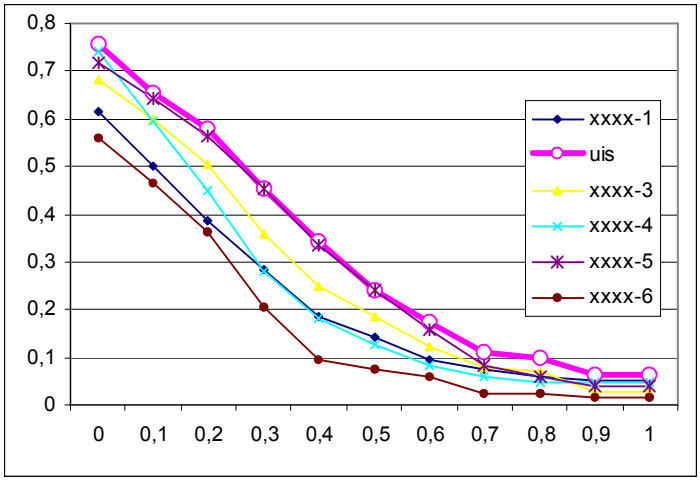


Рис. 4.6. Результаты дорожки РОМИП-2008 Legal adhoc (pd35).

В дорожке поиска по нормативно-правовой коллекции представленная модель показала лучший результат из 6 представленных алгоритмов, получив на первых 35 документах, которые были полностью оценены людьми-оценщиками, показатель средней точности MAP [227] – 29.6% (см. рис. 4.6), который превышает показатель следующего участника (27.6%) на 7%.

Чтобы проанализировать, насколько хорошо модель отработала на целевом множестве длинных информационных запросов, запросы были разбиты на несколько групп, одной из которых были длинные

информационные запросы, длиной более 3 слов, например, *уплата налога на прибыль организацией при отсутствии затрат* (27 запросов).

На основе этой классификации все оцененные запросы этой дорожки были разделены на соответствующие группы, для каждой группы отдельно была оценена средняя точность участников. На длинных информационных запросах была получена средняя точность MAP – 36%, что значительно превышает средний результат предложенного нами метода (29%), а также результат следующего участника (32%).

Проведенный анализ качества работы системы на разных группах запросов показывает, что важно уметь автоматически классифицировать поступающие запросы, и, в зависимости от класса запроса, применять несколько разные алгоритмы поиска.

4.3.4. Использование комбинированных моделей для поиска документов по запросам типа «формулировка проблемы» в правовой области

4.3.4.1. Особенность задачи

Несмотря на то, что подавляющее большинство запросов в поисковых системах относительно небольшой величины (в среднем 2-3 слова), существуют ситуации, когда пользователь задает достаточно длинный запрос. Необходимость в особенно длинных запросах возникает тогда, когда у пользователя есть какая-то проблема, и он обращается в интернет-форумы или вопросно-ответные сервисы, описывает свою проблему и ждет ответа от других пользователей форума или хотел бы найти документ, который помог бы ему справиться с его проблемой. При обращении в форум обязательным условием является то, что перед заданием вопроса людям, необходимо сделать усилия и попробовать найти ответ на свою проблему в предыдущих постах форума.

Задача поиска ответа на вопрос в виде формулировки проблемы значительно отличается от задач, решаемых в стандартных современных вопросно-ответных системах:

- количество запросов, похожих на вопросы, которые тестировались в рамках конференции TREC, достаточно мало.
- большинство вопросов представляет собой либо детальное описание ситуации и вопрос, специфичный для данной ситуации, либо совокупность структурно простых подвопросов, которые вместе также задают описание специфической правовой ситуации.
- при этом структурно сложные вопросы состоят из нескольких предложений и/или содержат несколько подвопросов.

При обработке структурно сложных вопросов имеются следующие сложности по сравнению с обработкой простых вопросов:

- автоматически трудно точно определить структуру вопроса – разбить его правильно на подвопросы, определить фокус вопроса;
- если часто можно ожидать, что ответ на структурно простой вопрос может содержаться в одном предложении текста, то ответ на структурно сложный вопрос может «собираться» из нескольких предложений документа.

В связи с этим для структурно сложных вопросов наиболее важным является поиск документов, содержащих описание соответствующей ситуации, при этом часто учет информации о структуре вопроса носит дополнительный характер.

Обработка длинных поисковых запросов в значительной степени отличается от обработки коротких поисковых запросов, которые являются наиболее распространенными запросами к поисковым системам.

Если при поиске по коротким запросам, поисковая система, скорее всего, найдет множество документов, включающих все слова запроса, и ее главной задачей является правильное упорядочение найденных документов, то при обработке длинных запросов к информационной системе в

подавляющем большинстве случаев не найдется ни одного документа или найдется всего несколько документов, содержащих все слова запроса. И, таким образом, основной задачей при обработке такого запроса является поиск и упорядочение документов, содержащих лишь часть слов запроса.

Казалось бы, векторные модели информационного поиска, которые описывают запрос и документы как вектора слов с весами, дают хороший базис для поиска ответов на длинные поисковые запросы, поскольку эта технология дает возможность установления частичного соответствия между запросом и документом.

Однако в реальности оказывается, что при использовании векторной модели часто поиск производится по относительно малозначащим словам запроса, в то время как очень важные слова запроса могут при сопоставлении исчезнуть. Как указывалось в п. 1.2.4.1, для того, чтобы в некоторой степени управлять формированием поискового запроса предлагается использование многошаговых булевских моделей. В следующем разделе будет описан алгоритм этого типа, который мы назвали «феноменологическая модель».

4.3.4.2. Алгоритм «Феноменологическая модель»

Феноменологическая модель – методика решения задачи поиска документов по запросу типа «формулировка проблемы» посредством моделирования понятиями ЛО содержания ситуации вопроса.

Феноменологическая модель преобразует запрос на естественном языке в булевский запрос вида конъюнкция дизъюнкций над понятиями ЛО:

$$\bigcap_i \bigcup_j c_{i,j},$$

где $c_{i,j} \in C$ — понятия лингвистической онтологии.

Элементами дизъюнкции могут быть понятия ЛО, которые рассматриваются как близкие по смыслу – они связаны между собой концептуальными путями определенного вида.

Действительно, вопрос пользователя не является последовательностью произвольных слов. В длинном вопросе многие упоминаемые понятия связаны между собой, например, принадлежат одной и той же области деятельности или одному и тому же типу.

Запрос типа «формулировка проблемы» описывает некоторую определенную ситуацию. Поэтому, чтобы иметь возможность дополнять булевское выражение понятиями из ЛО, необходимо иметь дополнительное подтверждение, что то или иное расширение подходит к описанной ситуации. Для этого используются так называемые информеры.

Информер – это совокупность наиболее характерных для данной поисковой выдачи понятий. В разных системах такого рода выдача может называться ассоциативный контекст, информационный портрет (см. например, [235] и др.). Понятия ЛО в информере упорядочиваются на основе веса, полученного по формуле типа tf.idf, когда частотность упоминания понятия в выдаче сопоставляется с частотностью упоминания понятия в коллекции.

В создаваемый булевский запрос могут быть добавлены понятия ЛО из дерева-вниз O^- или дерева-вверх O^+ одного из понятий запроса, если эти понятия входили в состав информера, т.е. принадлежали к множеству наиболее характерных понятий текущей выдачи. Дополнительное понятие вводится в дизъюнкцию к породившему его понятию запроса.

При этом феноменологическая модель рассматривается не как отдельная модель, а как отдельный компонент многошаговой модели. В частности, работа феноменологической модели начинается после предварительной работы векторной модели, которая отбирает 100 наиболее релевантных по запросу документов. Понятия ЛО из формулировки запроса упорядочиваются по количеству документов, найденных в этой выдаче – так определяются наиболее совместимые друг с другом понятия. Работа феноменологической модели начинается с наиболее частотного понятия в

упомянутой выдаче векторной модели, которое становится первым компонентом формируемого булевского выражения.

Рассмотрим работу феноменологической модели подробнее.

4.3.4.2.1. Обработка исходной формулировки вопроса

Работа модели начинается с того, что формулировка запроса сопоставляется с ЛО и составляется список понятий формулировки вопроса – C^q. Для многозначных слов проверяется, не разрешается ли многозначность на основе текущего списка понятий. Если есть возможность разрешить многозначность, то производится выбор значения или снятие пометки многозначности. Для каждого понятия формулировки определяется количество документов предварительной векторной выдачи, в которых оно встречается.

Следующее действие, которое нужно выполнить – построить списки близких по смыслу и поэтому потенциально объединяемых в дизъюнкции понятий запроса, на роль которых подходят понятия, связанные по иерархии отношений ЛО.

Между понятиями вопроса могут быть выявлены следующие типы взаимосвязей:

- понятия связаны между собой иерархическими путями P_{up} или P_{down} (Тип 1);
- понятия связаны между собой путем с перегибом вверх P_{updown} (Тип 2).

Пути типа 2 дополнительно классифицируются по длине и по типу входящих в перегиб отношений. Данная классификация связана с представлениями о близости понятий, не находящимися в непосредственном подчинении в иерархии ЛО. Типы перегибов упорядочены по предполагаемому снижению семантической близости между исходными понятиями.

Важной частью обработки формулировки запроса является формирование ядра запроса – C^{core}. Ядро запроса составляют понятия формулировки вопроса, для которых выполняются два условия:

- они порождаются по однозначным терминам или многозначность терминов была разрешена,
- их частота среди 100 документов, найденных по данному запросу по векторной модели, не менее 5.

Необходимость выделения ядра запроса связана с тем, что в запросе типа «формулировка проблемы» может быть большое количество случайно упомянутых понятий, в том числе, редко встречающихся в коллекции понятий. В таких случаях их относительно малая частотность в целевой коллекции не является критерием их важности для релевантной выдачи.

Остальные понятия формулировки вопроса также запоминаются для последующего уточнения запроса.

В ходе поиска документов нужно сформировать такой запрос к поисковой системе, чтобы он включал все понятия ядра для данной формулировки вопроса. В процессе формирования найденные документы складываются в копилку документов.

4.3.4.2.2. Построение формулы описания формулировки запроса

Формула описания запроса наращивается по шагам. Установлены следующие параметры алгоритма:

- doc_num_max – если число документов в выдаче меньше doc_num_max, то найденные на очередном шаге документы складываются в копилку документов (например, doc_num_max=50) в качестве потенциально релевантных;
- doc_num – если число документов в выдаче меньше этого числа, то запрос начинает расширяться, если больше – то сужаться (например, doc_num=20).

В каждой дизъюнкции формируемого булевского запроса D_i особо выделяется первый элемент, с которого начинается формирование данной дизъюнкции D_i^0. При добавлении понятий ЛО в создаваемый булевский запрос учитываются типы путей, между добавляемым понятием и начальными понятиями отдельных дизъюнкций D_i^0, уже входящих в запрос.

Построение формулы начинается с наиболее частотного в векторной выдаче понятия.

На каждом шаге выполняется сформированный запрос, оценивается количество найденных документов. Рассматриваются две основные ситуации: 1) больше ли количество документов в выдаче, чем doc_num или 2) меньше, чем doc_num.

В первом случае нужно запрос сужать, т.е. увеличивать конъюнкцию новыми элементами:

$$B \to (B \bigcap (c_i)),\ c_i \in C^{core},\ c_i \notin C^B$$

где C^{core} – понятия из ядра вопроса, C^B– понятия, уже включенные в формируемый булевский запрос B.

В качестве нового элемента конъюнкции берется понятие из ядра формулировки ядра запроса, не связанное или с наименьшим весом связанное по ЛО с начальными понятиями дизъюнкций cd_i^0 текущего булевского выражения. Тем самым более близкие понятия оставляются как ресурс для возможного расширения запроса. Это дает возможность одни и те же понятия в некоторых запросах располагать в разных элементах конъюнкции (т.е. использовать для сужения запроса), а в других – как элементы одной и той же дизъюнкции (использовать для расширения запроса). Если таких (наиболее далеких) понятий несколько, то выбирается первое по списку понятий-кандидатов на добавление.

Во втором случае необходимо расширять формируемый запрос, дополняя дизъюнкции.

$$B = B^{k-1} \bigcap (\mathrm{c}d_0^{\mathrm{k}} \bigcup \mathrm{c}d_1^{\mathrm{k}} ...) \rightarrow (B^{k-1} \bigcap (\mathrm{c}d_0^{\mathrm{k}} \bigcup \mathrm{c}d_1^{\mathrm{k}} .. \bigcup c_i))$$

В качестве понятий, которыми могут быть дополнены дизъюнкции, могут использоваться:

- понятие c_i формулировки вопроса, еще не включенное в формируемый булевский запрос ($c_i \in C^{core}$, $c_i \notin C^B$) и имеющее разрешенные пути отношений к начальным понятиям дизъюнкций cd_i^0;
- понятие c_i, которого нет в формулировке запроса, но которое находится в дереве-вверх или в дереве-вниз начальных понятий дизъюнкций cd_i^0 и которое входит в совокупность понятий информера последнего выполненного запроса C^{inf}, как наиболее характерное для последней выдачи документов, тем самым планируемое расширение запроса подтверждается текущей выдачей.

Если таких понятий не имеется и есть еще понятия ядра формулировки, которые не включены в булевский запрос, то последняя дизъюнкция запроса начинает наращиваться этими оставшимися понятиями.

$$B = B^{i-1} \bigcap (\mathrm{D}_0^{\mathrm{k}} \bigcup \mathrm{D}_1^{\mathrm{k}} ...) \rightarrow (B^{i-1} \bigcap (\mathrm{c}d_0^{\mathrm{k}} \bigcup \mathrm{c}d_1^{\mathrm{k}} .. \bigcup c_i)),$$
$$c_i \in C^{core}, c_i \notin C^B,$$

Результат исполнения последнего запроса (который содержит все понятия ядра) заносится в копилку. Отметим, что операции сужения и расширения запроса всегда применимы, пока не все понятия ядра вопроса включены в формулу. Таким образом, алгоритм гарантирует включение всех понятий ядра вопроса в формулу. Документы, полученные работой алгоритма, присоединяются к документам, полученным векторной моделью и направляются на дальнейший анализ, который производится подобно

процедуре, описанной в п. 4.3.3, посредством оценки наиболее наполненных элементами запроса и расширением запроса предложений

Приведем пример сформированного феноменологической моделью булевского запроса для следующей формулировки запроса (используется Общественно-политический тезаурус):

> Вопрос: *Туристическая фирма (турагент) занимается реализацией путевок сторонних организаций в санаторно-курортные и оздоровительные учреждения. В соответствии с действующим законодательством реализация такого продукта не подлежит обложению НДС. Однако в ходе проверки налоговой инспекцией нам были предъявлены санкции за неуплату налога с суммы агентского вознаграждения. Правы ли в данном случае налоговые органы? ("Консультант бухгалтера", N 7, июль 2001 г.)*

Для данной формулировки выделены следующие понятия ядра, которые необходимо «уложить» в булевское выражение (перечислены по алфавиту):

АГЕНТСКОЕ ВОЗНАГРАЖДЕНИЕ
НАЛОГ НА ДОБАВЛЕННУЮ СТОИМОСТЬ
НАЛОГОВАЯ СЛУЖБА
НАЛОГОВОЕ ОСВОБОЖДЕНИЕ
ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ПУТЕВКИ НА ОТДЫХ И ЛЕЧЕНИЕ
САНАТОРНО-КУРОРТНОЕ ЛЕЧЕНИЕ
СТОРОННЯЯ ОРГАНИЗАЦИЯ
ТУРАГЕНТ
ТУРИСТИЧЕСКАЯ ФИРМА

Формирование булевского выражения началось с понятия *ТУРАГЕНТ*. По данному запросу в коллекции найдено 66 документов, что больше

установленного параметра doc_num=20, поэтому в конъюнкцию добавляется понятие *ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*, что приводит к величине выдачи 8 документов. Запрос необходимо расширять. Из формулировки извлекается понятие *ТУРИСТИЧЕСКАЯ ФИРМА*, являющееся вышестоящим понятием для понятия *ТУРАГЕНТ (*см. рис. 4.7), и вносится в соответствующую дизъюнкцию, получается такой запрос:

(ТУРАГЕНТ OR ТУРИСТИЧЕСКАЯ ФИРМА)
AND
(ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ)

В результате выполнения такого запроса находится 16 документов. Запрос необходимо расширять дальше. Такую возможность дает информер, сформированный по последнему булевскому запросу. На седьмом месте самых характерных понятий для данной выдачи находится понятие *САНАТОРИЙ*, который является видом понятия *ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*, и, таким образом, пополняется соответствующая дизъюнкция. Получается следующий булевский запрос:

(*ТУРАГЕНТ OR ТУРИСТИЧЕСКАЯ ФИРМА*)
AND
(*ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ OR САНАТОРИЙ*)

Выдача данного запроса содержит 22 документа, и запрос опять можно уточнять.

В результате последовательности шагов работы алгоритма было сформировано следующее булевское выражение (жирным выделены начальные понятия каждой дизъюнкции):

*(**ТУРАГЕНТ*** *OR*
ТУРИСТИЧЕСКАЯ ФИРМА *OR*
ТУРИСТИЧЕСКИЙ СЕРВИС *OR*
ПОСРЕДНИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ *OR*
АГЕНТСКОЕ ВОЗНАГРАЖДЕНИЕ *OR*

ПОСРЕДНИЧЕСКАЯ ОРГАНИЗАЦИЯ OR
ПУТЕВКИ НА ОТДЫХ И ЛЕЧЕНИЕ)

AND

***(ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ** OR*
САНАТОРИЙ OR
ДОМ ОТДЫХА OR
ОТДЫХ OR
ПРОФИЛАКТОРИЙ OR
ДЕТСКОЕ ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ OR
СТОРОННЯЯ ОРГАНИЗАЦИЯ)

AND

***(САНАТОРНО-КУРОРТНОЕ ЛЕЧЕНИЕ** OR*
САНАТОРНО-КУРОРТНАЯ ПУТЕВКА OR
ЗДРАВООХРАНЕНИЕ OR
ЛЕЧЕНИЕ)

AND

(НАЛОГ НА ДОБАВЛЕННУЮ СТОИМОСТЬ)

AND

***(НАЛОГОВОЕ ОСВОБОЖДЕНИЕ** OR*
НАЛОГОВАЯ СЛУЖБА)

По этому запросу был найден 51 документ.

Помимо понятий Общественно-политического тезауруса**,** найденных в исходной формулировке, феноменологическая модель добавила в булевское выражение следующие понятия:

- *ТУРИСТИЧЕСКИЙ СЕРВИС*,
- *ПОСРЕДНИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ*,
- *ПОСРЕДНИЧЕСКАЯ ОРГАНИЗАЦИЯ*,
- *САНАТОРИЙ*,
- *ДОМ ОТДЫХА*,
- *ОТДЫХ*,
- *ПРОФИЛАКТОРИЙ*
- *ДЕТСКОЕ ОЗДОРОВИТЕЛЬНОЕ УЧРЕЖДЕНИЕ*

- *САНАТОРНО-КУРОРТНАЯ ПУТЕВКА,*
- *ЗДРАВООХРАНЕНИЕ,*
- *ЛЕЧЕНИЕ*

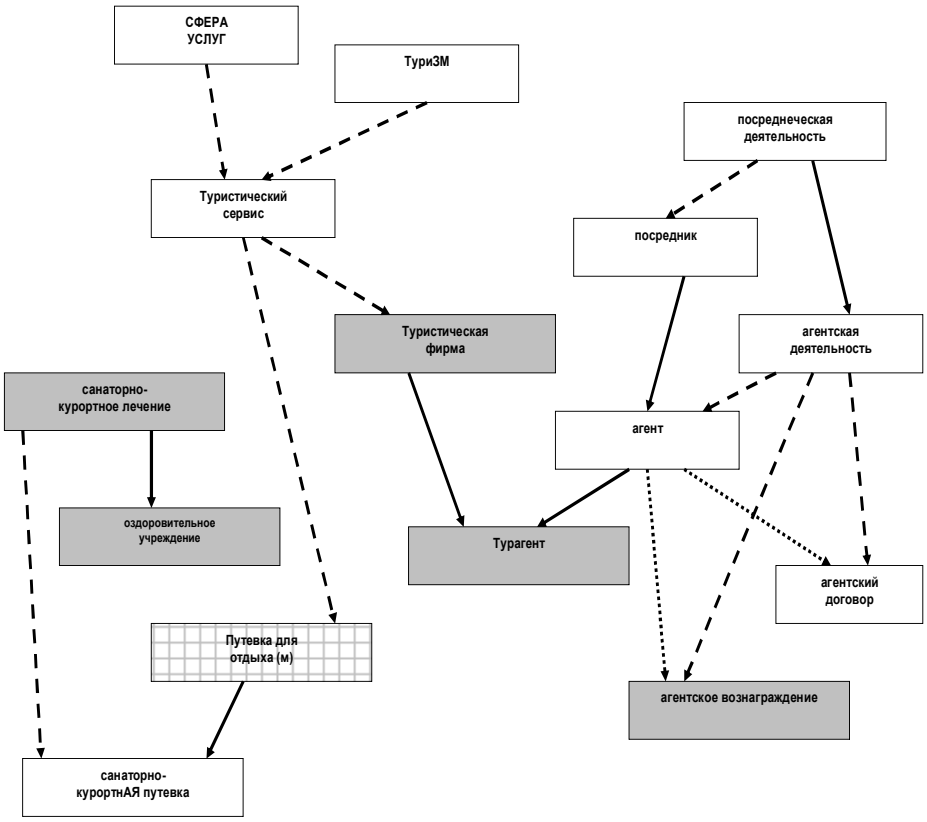


Рис. 4.7. Фрагмент Общественно-политического тезауруса, использованный для обработки вопроса о налогообложении туристической фирмы

На рис. 4.7 показан "туристический" фрагмент Общественно-политического тезауруса, относящиеся к теме вопроса. Серые квадратики обозначают понятия, которые исходно были обнаружены в формулировке вопроса.

4.3.4.2.3. Применение феноменологической модели

Знания, описанные в онтологии, могут быть неполными, и в очередной формулировке запроса могут потребоваться знания, не отраженные в ЛО. Поэтому феноменологическая модель не применяется отдельно, а входит в состав многошаговой модели, описанной в п. 4.3.3.

Феноменологическая модель работает после комбинированной векторной модели. Найденные в формулировке понятия ЛО упорядочиваются по количеству документов, в которых они упоминаются в этих 100 документах для работы феноменологической модели. Таким образом, предполагается, что булевские запросы феноменологической модели будут строиться на понятиях ЛО, которые наиболее часто упоминаются в связи друг с другом.

В результате работы модели и исполнения построенных булевских запросов «копилка» документов для дальнейшего анализа пополняется дополнительными документами. Кроме того, в процессе своей работы феноменологическая модель расширяет запрос понятиями ЛО, которые не были упомянуты в запросе, и эти дополнительные понятия будут также придавать дополнительный вес найденным документам.

Суть дальнейшего анализа документов заключается в том, чтобы дополнительно проанализировать все найденные на предыдущих этапах документы (100 документов от смешенной векторной модели и 30-100 документов от феноменологической модели). Наилучшими считаются документы, в которых максимальное число найденных элементов запроса, найдено в 2 парах соседних предложений документа (см. п. 4.3.3).

Формула предложения (4.7) дополняется еще и весом понятий ЛО, которые не были упомянуты в формулировке запроса, но были получены в процессе расширения по феноменологической модели. Таким образом, вес отдельного предложения вычисляется следующим образом по сравнению с формулой (4.8):

$$W_s = \alpha_2 \Sigma w_{wordi} + (1 - \alpha_2) \Sigma w_{concj} + \alpha_3 \Sigma w_{exp} \quad (4.9)$$

где w_{wordi}, w_{concj} – веса слов и понятий из исходной формулировки, w_{exp} – это вес понятия ЛО, которого не было в исходной формулировке, но который был добавлен в расширенный запрос на этапе работы феноменологической модели:

$$w_{exp}(c_i) = idf(c_i)$$

«Усиленный» вес за счет дополнительных предложений считается так, как описано в п. 4.3.4, в дополнительных предложениях также учитываются дополнительно полученные понятия ЛО. Полученный вес предложения замешивается с исходным весом предложения, полученным по векторной модели первого этапа.

Таким образом, выполнение феноменологической модели дает возможность привлечь дополнительное число документов для последующего анализа, и, кроме того, учесть вес понятий, полученных как расширение булевского запроса.

4.3.4.2.4. Оценка качества работы комбинированной модели

Качество комбинированной модели, включая феноменологическую модель, тестировалось на 165 запросах типа «формулировка проблемы» в юридической области экспертами-юристами на коллекции документов, отвечающих на такие вопросы (40 тысяч документов). Оценка производилась по показателю точности по первым пяти документам – precision (5). В таблице 4.3 приводится точность для разных методов по 5, 10, 15, 20 документам. Таким образом, достигнутый показатель точности precision(5) по комбинированной модели, включающей феноменологическую модель, составляет 76.48%. Учет дополнительных знаний привел к росту показателя precision(5) на более чем 12% по сравнению с наилучшими результатами поиска, в котором не использовалась информация из ЛО и феноменологическая модель.

Табл. 4.3. Точность по 5 первым документам по различным применяемым моделям.

Метод	5	10	15	20
Поиск по комбинированной векторной модели	**55.88**	49.45	44.28	40.55
Максимальный результат, полученный по леммам (векторная модель + упорядочение по предложениям + замешивание полученных весов) – то есть без всякого участия тезауруса	**68.00**	57.70	52.89	47.70
Комбинированная модель, включая феноменологическую модель	**76.48**	65.21	57.54	51.82
Яндекс-сервер (лучший результат по настройкам)	50.84	43.07	38.85	35.37

4.4. Лингвистическая онтология как ресурс для автоматической рубрикации текстов

4.4.1. Технология автоматического рубрицирования на основе ЛО

Задачей автоматической рубрикации (классификации) текстов является автоматическое отнесение к одной категории из заранее заданной системы категорий (рубрикатора), т.е. необходимо построить функцию отображения

$$\Phi: \; D \times C \to \{0, 1\}$$

где D – это множество документов, C – множество заданных категорий.

Как известно, существуют два основных подхода к автоматическому рубрицированию документов – подход, основанный на знаниях, когда эксперты вручную создают описания рубрик, и подход на основе машинного обучения, при котором модель рубрикации создается на основе уже отрубрицированных документов. Известно, что одной из проблем

инженерного подхода является сложность составления описания конкретной рубрики. Развитая ЛО позволяет в значительной мере смягчить эту проблему.

Предлагается метод автоматического рубрицирования, который основывается на трех основных положениях:

- содержание рубрики описывается как булевский запрос над понятиями ЛО,

- текущий рубрикатор связывается с ЛО посредством небольшого числа опорных понятий, рубрики остальных понятий ЛО выводятся по связям внутри ЛО, тем самым при описании очередного рубрикатора используется большой объем накопленных в лингвистической онтологии знаний.

- процедура рубрикации базируется на автоматически построенном тематическом представлении документов, которое моделирует основную тему и подтемы документа наборами (тематическими узлами) близких по смыслу понятий, упомянутых в документе [230, 251, 256].

Новым в данном методе автоматической рубрикации является существенно использование свойств отношений в ЛО и логического вывода на отношениях. Кроме того, новизна метода заключается в том, что базой для проставления весов рубрик является тематическое представление текстов. Это позволяет при выводе рубрик для документа использовать выявленные отношения между понятиями текста, а также такая основа рубрикации дает возможность обрабатывать тексты разных типов и размеров: нормативные акты, газетные статьи, новостные сообщения, научные публикации в области гуманитарных наук, социологические опросы.

В следующих подразделах предложенная технология автоматической рубрикации рассматривается более подробно.

4.4.2. Описание смысла рубрики понятиями ЛО

При создании лингвистического профиля рубрикатора каждая рубрика R описывается дизъюнкцией альтернатив, каждый элемент дизъюнкции представляет собой конъюнкцию:

$$R = \bigcup_i D_i \; ; \qquad D_i = \bigcap_j K_{ij} \; , \tag{4.9}$$

Элементы конъюнкции в свою очередь описываются экспертами с помощью так называемых «опорных» понятий ЛО. Для каждого опорного понятия задается правило его расширения $f(\cdot)$, определяющее, каким образом вместе с опорным понятием учитывать подчиненные ему по иерархии понятия: без расширения (обозначается символом «N»), полное расширение по дереву иерархии онтологии (символ «E»), расширение только по родовидовым связям (символ «L»), расширение по всем видам отношений на один уровень иерархии (символ «W»), расширение на один уровень иерархии, не включая отношения НИЖЕ (символ «V»).

В результате расширения элемент конъюнкции преобразуется в дизъюнкцию понятий ЛО: $K_{ij} = \bigcup_m c_{ijm}$

Опорное понятие может быть как «положительным», т.е. добавлять ниже расположенные понятия в описание элемента конъюнкции, так и «отрицательным», то есть вырезать из описания рубрики свои подчиненные понятия. Последовательность учета положительных и отрицательных опорных понятий регулируется заданием специального атрибута. Результатом применения расширения опорных понятий является совокупность понятий ЛО, полностью описывающая элемент конъюнкции:

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn}) = \bigcup_k d_{ijk} \; . \tag{4.10}$$

Рассмотрим фрагмент представления рубрики 200.020.020 «Встречи на высшем уровне» из Классификатора правовых актов РФ ([311] - более 1000

рубрик). Языковые выражения, записанные курсивом, выводятся на основе исходного описания рубрики автоматически (рис. 4.8):

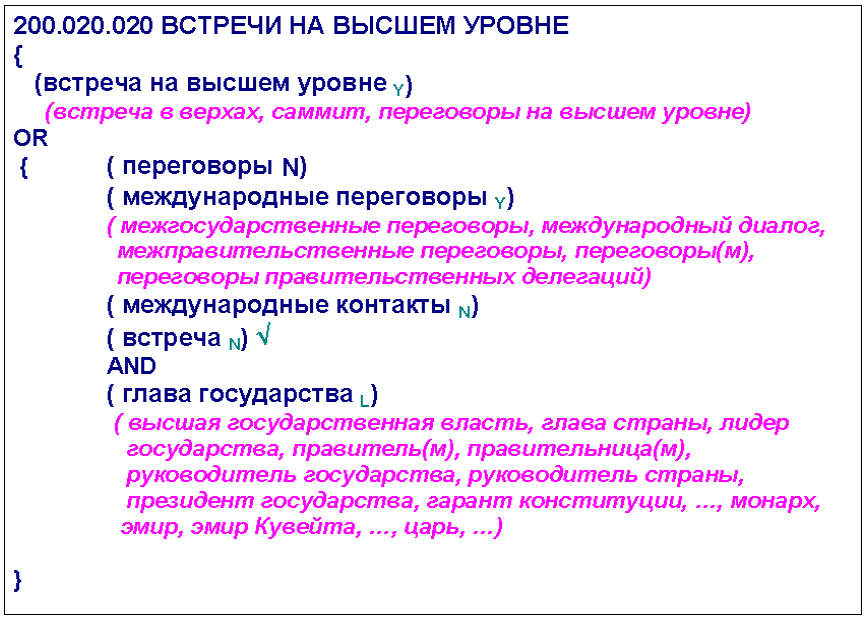


Рис.4.8. Расширенное представление рубрики понятиями ЛО

4.4.3. Автоматическое рубрицирование на основе тематического представления

Как отмечалось в предыдущем разделе, рубрика представляется в виде логического условия над понятиями ЛО:

$$R = \bigcup_i D_i = \bigcup_i \left[\bigcap_j K_{ij} \right] = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right]. \qquad (4.11)$$

Таким образом, оценка релевантности содержания текста рубрике (вес рубрики) может быть рассчитана на основе информации о весах понятий в тексте, входящих в ее описание.

Вес элемента конъюнкции рассчитывается по формуле:

$$\theta(K_{ij}) = \max_k \left(\theta(d_{ijk}) \right), \qquad (4.12)$$

где d_{ijk} – понятия ЛО, полученные из опорных понятий, приписанных элементу конъюнкции экспертом, посредством применения функции расширения; $\theta(d_{ijk})$ – вес понятия, полученный на основе построенного тематического представления.

Вес конъюнкции в целом предназначен учитывать не только сумму весов составляющих его элементов, но и меру их близости в тексте:

$$\theta(D_i) = \frac{\sum_{j=1}^{m}\theta(K_{ij}) + \sum_{j<k} S(K_{ij}, K_{ik})}{m + C_m^2}, \qquad (4.13)$$

здесь $S(K_{ij}, K_{ik}) = \min\{1.0; \frac{\sum s(c_{ijq} \in K_{ij}, d_{ikw} \in K_{ij})}{\max s(c \in D, d \in D)}\}$

- сумма всех текстовых связей между понятиями одного элемента конъюнкции и понятиями другого, деленная на значение максимальной текстовой связи между любыми двумя понятиями текста. Этот член равен обычно единице для сильно связанных конъюнктов и принимает малое значение, если понятия различных конъюнктов обсуждались в разных местах текста.

Вес рубрики представляет собой максимум весов входящих в описание рубрики альтернатив. В случае имеющихся иерархических связей между рубриками оценка релевантности нижестоящих рубрик переносится на вышестоящие. Так что при запросе по вышестоящей рубрике будут выходить и документы, к которым были приписаны нижестоящие рубрики.

Алгоритм рубрицирования работает следующим образом. Для всех понятий ЛО, найденных в тексте, определяется множество рубрик, которые могут быть определены в тексте. Для каждой рубрики происходит расчет ее веса по формулам (4.12) и (4.13). В результирующем множестве остаются рубрики, вес которых превосходит задаваемый заранее для коллекции порог.

Применение описанной технологии для нескольких систем рубрикации для различных текстовых коллекций показали, что описание рубрикатора посредством опорных понятий служит и как основа для соответствующих организационных решений:

- является прообразом свободного от субъективизма комментария к рубрикатору, который может пополняться и уточняться;
- при выводе рубрики всегда можно показать/объяснить, почему была выведена та или иная рубрика, что позволяет быстро уточнять описание рубрик, анализируя замеченные ошибки рубрикации.

4.4.4. Эксперимент по автоматической рубрикации текстов в рамках семинара РОМИП 2007

Опишем результаты работы системы автоматического рубрицирования, основанной на одной из реализаций модели ЛО – Общественно-политическом тезаурусе – в задаче классификации Web-страниц в рамках семинара РОМИП 2007 [227]. Исходный набор данных включал в себя коллекцию страниц с сайтов белорусского интернета BY.web и коллекцию DMOZ, используемую в качестве обучающего множества. Обучающее множество состоит из сайтов, но не обязательно все страницы сайта относятся к одной теме. Рубрикация должна была быть выполнена для 247 рубрик рубрикатора DMOZ.

При выполнении данного эксперимента была поставлена задача выяснения, сколько времени нужно потратить на описание заданных рубрик с использованием из информации из тезауруса, и каких показателей качества рубрицирования можно достигнуть.

Работа по описанию 247 рубрик задания была выполнена за 8 часов рабочего времени. В опорных булевских выражениях было использовано около 900 понятий Общественно-политического тезауруса (в качестве базовой ЛО), в расширенных булевских выражениях содержится около 40

тысяч понятий (с повторениями). Каждому понятию тезауруса соответствует в среднем два-три языковых выражения (слова или словосочетания).

На выбранных организаторами для тестирования 19 рубриках система рубрикации показала наивысшие показатели классификации по F-мере (см. п. 1.2.5.1). По метрике AND, считающей релеватными рубрики документы с учетом мнений обоих оценщиков, для оцененных документов величина F-меры составила 44%, что почти на 42% превышает результаты следующей по величине F-меры системы рубрикации (31%). По метрике OR (документ считается относящимся к рубрике, если хотя бы один из асессоров отнес его к данной рубрике) для оцененных документов величина F-меры составила 72%, что более чем на 56% превысило показатели следующей по качеству результатов системы (46%).

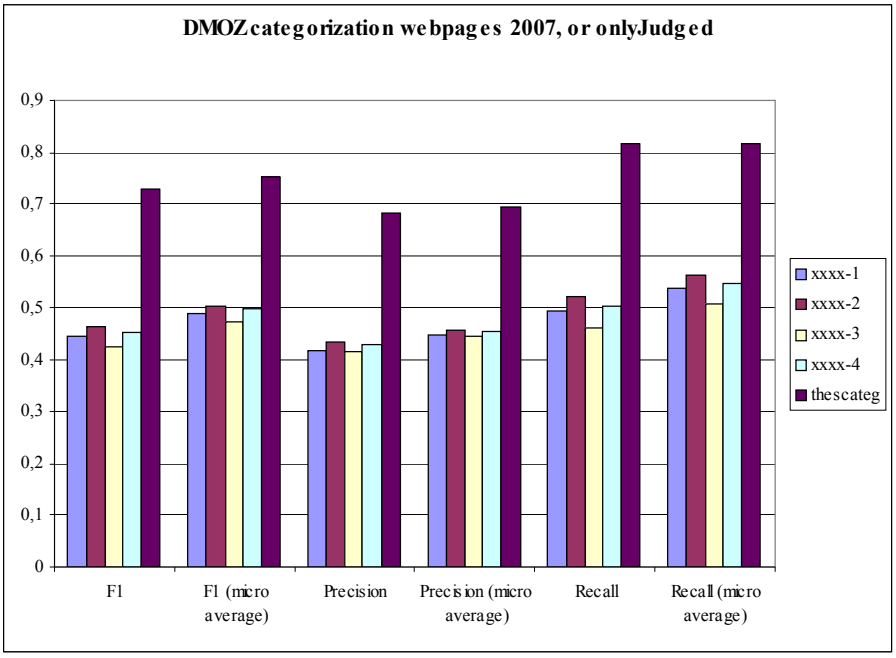


Рис. 4.9. Результаты рубрикации веб-страниц на основе тезауруса РуТез в экспериментах РОМИП-2007

На Рис. 4.9 приведены результаты для дорожки классификации Веб-страниц коллекции ROMIP.BY. Достижение показателей качества

рубрицирования при нестрогом согласии между экспертами (полнота = 81.7%; точность = 68.2%; F-мера = 72.9%) следует признать весьма успешным для 8 часов трудозатрат экспертов.

Из информации на семинаре РОМИП-2007 известно, что другие методы представляли собой модификации метода машинного обучения SVM – одного из самых успешных методов, применяемых в автоматической рубрикации текстов. Анализ данных коллекции показал, что проблемы методов машинного обучения связаны со значительной противоречивостью разметки в коллекции, выданной для обучения.

В текущем эксперименте у нас не было возможности сделать предварительный прогон и исправить ошибки и неточности описания. Поэтому имеется очевидная возможность улучшения полученных результатов тематической классификации веб-страниц на основе знаний, описанных в ЛО.

Таким образом, при подходе, основанном на знаниях, накопленная в ЛО информация дает возможность более быстрого и качественного описания содержания рубрик рубрикатора. Продемонстрированные в экспериментах результаты показывают, что некоторую значимую часть знаний о современной жизни общества и современном языке деловой прозы нам удалось описать и упорядочить в рамках понятийных структур такой лингвистической онтологии как Общественно-политический тезаурус.

На основе данного метода автоматической рубрикации было создано более 20 систем автоматической рубрикации документов.

4.5. Методы автоматического аннотирования текстов на основе лингвистической онтологии

Тематическое представление одного или группы тематически связанных документов, построенное на основе лингвистических онтологий, позволяет развивать методы, на основе которых можно лучше решать типичные проблемы автоматического аннотирования, а именно,

обеспечивать полноту представления информации, снижать повторы, обеспечивать связность и понятность аннотации.

В данном разделе описывается метод автоматического аннотирования одного и многих документов, который базируется на тематическом представлении текстов. Новизна метода состоит в том, что метод использует предположение об иерархической структуре связного текста и пропозиции как основного структурного элемента этой иерархии.

4.5.1. Метод автоматического аннотирования отдельного текста на основе тематического представления

При построении тематического представления текста в виде совокупностей близких по смыслу понятий, упоминаемых в тексте (тематических узлов), выявляются основных участников ситуации, описываемой в тексте. Так называемые основные тематические узлы моделируют главных участников описываемой ситуации. Суть текста составляет описание взаимодействия между главными участниками.

Таким образом, то новое и важное, что несет в себе текст и что должна отразить в себе аннотация, это именно то, каким образом взаимодействуют между собой эти главные участники. Отсюда следует первый принцип составления аннотаций: важными (информативными) и, следовательно, возможно включенными в аннотацию считаются те предложения текста, которые содержат, по крайней мере, два понятия, входящих в состав разных основных тематических узлов текста (рис. 4.10). Напомним, что алгоритмы автоматического аннотирования на основе лексических цепочек и WordNet при извлечении предложений требуют присутствия одного элемента из основных лексических цепочек (см. п. 3.3) [277, 235].

Предложений, содержащих понятия одних и тех же двух основных тематических узлов, в тексте может оказаться достаточно много. Для аннотации необходимо выделить одно предложение, в котором

взаимодействие этих двух основных тематических узлов характеризуется “наилучшим образом”.

Не все основные участники начинают обсуждаться в тексте сразу, с первого предложения – часть из них возникает в последующих предложениях. Чтобы сохранить связность и последовательность изложения текста, автор именно в этом первом предложении новой темы должен наиболее точно указать связь новой темы со всем предшествующим текстом. Таким образом, следуя за автором при вводе нового тематического элемента, можно повысить общую связность аннотации, т.е. второй принцип составления аннотации отдельного документа состоит в том, что для каждой пары выявленных основных тематических элементов текста (основных тематических узлов) в аннотацию выбираются предложения, содержащие первое вхождение этой пары, следуя по порядку текста.

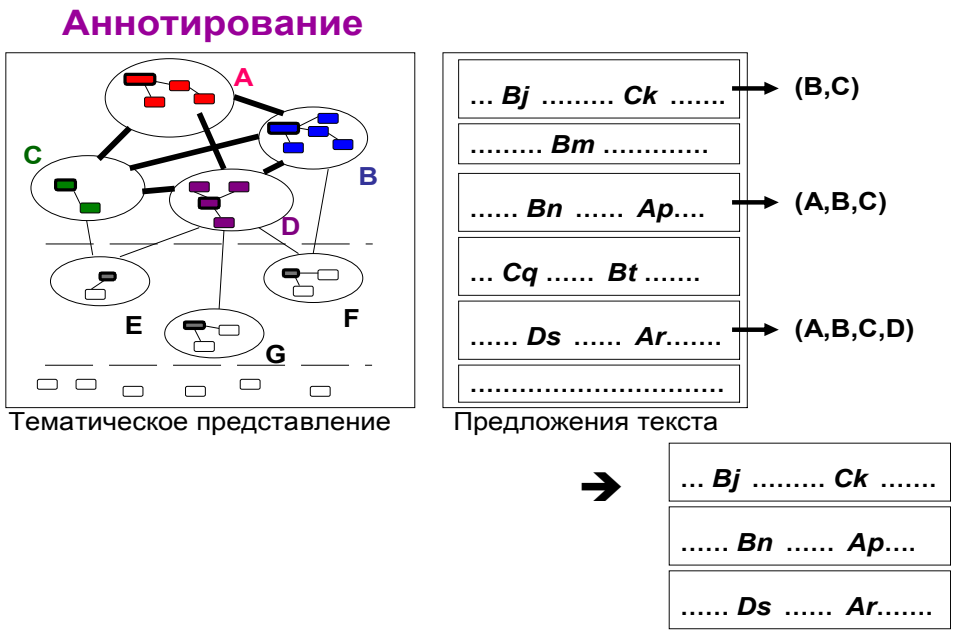


Рис. 4.10. Учет тематического представления при формировании аннотации.

Нужно отметить, что при хорошем покрытии предметной области знаниями, описанными в ЛО, появление в очередном предложении новой

темы выявляется весьма точно, а это означает, что связность получаемой аннотации в среднем весьма высока.

Построение аннотации реализуется следующим образом:

1) Для построения аннотаций сначала формируется множество "аннотационных" предложений-кандидатов, которые не являются вопросительными или восклицательными предложениями.

2) Перед построением аннотации создается таблица всех возможных пар основных тематических узлов $(tnode_i, tnode_j)$, $tnode_i, tnode_j \in Tnode^M$, и устанавливается соответствие между предложениями текста и данными парами основных тематических узлов. Отношение *rt* между предложением текста s_k и парой тематических узлов $(tnode_i, tnode_j)$ устанавливается, если в этом предложении упоминаются два различных понятия c_m и c_n такие, что $c_m \in tnode_i$, $c_n \in tnode_j$.

3) Начиная с начала текста, отбираются такие предложения s_k, которые содержат еще не упоминавшуюся в аннотации пару разных тематических узлов, т.е. $rt(s_k, (tnode_i, tnode_j))$ *и не установлено отношение* $rt(s_l, (tnode_i, tnode_j))$, *где* s_l – одно из предшествующих предложений текста T, $l<k$.

Серьезной проблемой автоматического аннотирования является проблема местоимений, которые могут появиться в выбранных предложениях и служить ссылкой на такие предложения текста, которые не вошли в состав аннотации.

В настоящее время в случаях, когда очередное предложение текста подходит для аннотации, но содержит местоимение, принимается одно из следующих решений:

1) если предыдущее предложение входит в состав аннотации, то и данное предложение включается в состав аннотации;
2) если предыдущее предложение не входит в состав аннотации, то проверяется, нельзя ли это предыдущее предложение включить в

состав аннотации. Для этого необходимо, чтобы оно не содержало местоимений или следовало за предложением, включенным в аннотацию.

3) в остальных случаях предложение с местоимением не включается в состав аннотации.

Качество предложенного метода технологии автоматического аннотирования тестировалось на конференции SUMMAC (summarization conference) [126, 127]. В качестве лингвистической онтологии, программа автоматического аннотирования использовала английский перевод Общественно-политического тезауруса.

Задача, в рамках которой тестировался изложенный метод автоматического аннотирования, состояла в следующем. Каждый участник соревнования получал на две недели 1000 документов и должен был представить две аннотации – аннотацию наилучшей длины (т.е. система сама определяла длину аннотации) и 10-процентную аннотацию, т.е. аннотацию, составляющую 10 процентов длины исходного текста.

Тестирование в процессе соревнования относилось к так называемому классу внешних тестирований (extrinsic), то есть проверялось, насколько порожденная аннотация пригодна для решения некоторой внешней задачи.

Внешней задачей в данном случае была задача рубрикации. Все документы, выданные для обработки, относились к двум большим темам «Мировая экономика» и «Налоги». При этом по полному тексту документа, его можно было отнести к более подробным рубрикам. Так, например, для рубрики «Мировая экономика» такими подрубриками были:

- экспорт в промышленности,
- внешняя торговля,
- международная борьба с наркотиками,
- иностранные производители автомобилей.Таким образом, если аннотация сделана правильно и сохраняет основную тему документа, люди-оценщики должны отнести аннотацию документа к той же подрубрике, что и

сам документ. При этом каждому человеку-оценщику давался документ, который мог оказаться аннотацией, начальным фрагментов документа или полным текстом. По ошибкам отнесения можно было оценить качество полученной аннотации.

Качество рубрикации документов по аннотации, и таким образом, собственно аннотаций оценивалось по стандартным метрикам, использующимся при оценивании систем автоматической рубрикации: точность, полнота и F-мера. Представленная нами система имела лучший показатель F-меры для аннотаций наилучшей длины и показатели 10-процентных аннотаций были лучше, чем средние [127] (рис. 4.11). По оси X отражается время, за которой человек принимал решение на основе аннотации, по оси Y – правильность проставленных аннотации рубрик.

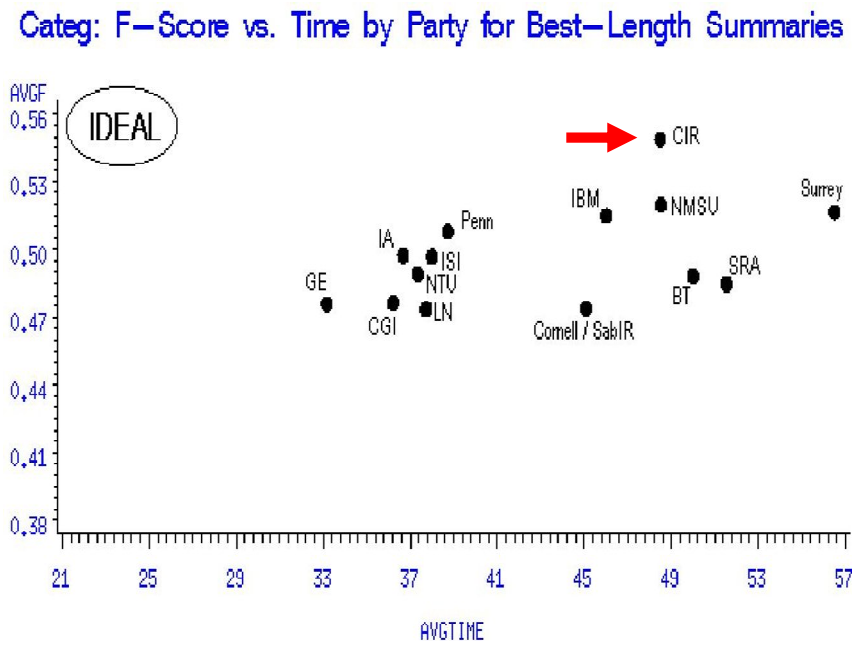


Рис. 4.11. График отражает соотношение качества аннотации и времени на принятие решения. Стрелкой показан результат описанного алгоритма.

В качестве примера работы описанного метода автоматического аннотирования рассмотрим следующий текст:

Китай и Тайвань установили авиасообщение

после 60-летнего перерыва

После почти 60-летнего перерыва открылось регулярное авиасообщение между Тайванем и материковым Китаем. Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу, передает «Би-би-си». Ожидается, что аэропорты острова будут принимать рейсы из пяти китайских городов: Пекина, Шанхая, Гуанчжоу, Сямэня и Нанкина. Договоренность о прямых регулярных авиарейсах была достигнута в середине июня 2008 года на переговорах между руководством Тайваня и Китая. Восстановление авиасообщения произошло не в последнюю очередь благодаря победе на выборах главы администрации Тайваня в марте 2008 года сторонников тесного сотрудничества с материковым Китаем. Прямых регулярных авиарейсов между Тайванем и Китаем не осуществлялось с 1949 года, когда Тайвань стал убежищем потерпевших поражение в гражданской войне с коммунистами сторонников партии Гоминьдан. До июля 2008 года прямые рейсы между материковым Китаем и Тайванем осуществлялись только по спецдоговоренности, в основном – в дни праздников, напоминает Лента.ру

Приведем примеры тематических узлов, созданных в процессе обработки этого текста (центр тематического узла выделен сдвигом влево; указана также частота упоминания понятия в тексте):

КИТАЙ	*8*
ГУАНЧЖОУ	*2*
ШАНХАЙ	*2*
НАНКИН	*1*
ПЕКИН	*1*
ТАЙВАНЬ	*7*
ТАЙБЕЙ	*1*
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	*2*
АВИАРЕЙС	*1*
АЭРОПОРТ	*1*
ПОЛИТИЧЕСКАЯ ПАРТИЯ	*1*
КОММУНИСТ	*1*
ПРАВИТЕЛЬСТВО	*1*
ПУБЛИЧНАЯ ВЛАСТЬ	*1*

В рассматриваемом примере тематического представления основными тематическими узлами стали узлы с главными дескрипторами *КИТАЙ (1)*,

ТАЙВАНЬ (2), АВИАЦИОННЫЕ ПЕРЕВОЗКИ (3), ГОРОД (4), РЕЙС (5), БИ-БИ-СИ (6).

Для текста примера получаем следующую аннотацию, в которой упомянуты все основные тематические узлы данного документа (представители основных тематических узлов выделены в тексте, указаны номера узлов):

***Китай*$_1$** *и* **Тайвань$_2$** *установили* ***авиасообщение*$_3$** *после 60-летнего перерыва*

Первый чартерный ***рейс*$_5$** *с 250 пассажирами уже прибыл в столицу* ***Тайваня*$_2$** *из* ***китайского*$_1$ *города*$_4$ *Гуанчжоу*$_1$,** *передает* ***Би-би-си*$_6$**. *Ожидается, что* ***аэропорты*$_3$** *острова будут принимать рейсы из пяти* ***китайских городов*$_4$ *Пекина*$_1$, *Шанхая*$_1$, *Гуанчжоу*$_1$,** *Сямэня и* ***Нанкина*$_1$**.

Отметим, что в аннотации пропущено первое предложение, которое не содержит новой пары тематических узлов по сравнению с заголовком текста.

4.5.2. Построение структурной тематической аннотации текста

Для некоторых типов текстов хорошая (связная и понятная) аннотация может быть построена не всегда [108, 279]:

- большинство таких документов как законы, президентские и правительственные документы, международные договоры имеют очень сложную структуру, поэтому часто аннотация, основанная на любых принципах выбора информативных фрагментов, может быть слишком длинна, тяжеловесна и неясна;
- автоматические аннотации газетных интервью обычно обрывочны и несвязны;
- в ограниченные по длине аннотации текстов больших размеров могут не уместиться важные темы текста;

- и, наконец, аннотации на исходном языке могут оказаться бесполезными для пользователей многоязычных информационно-поисковых систем. Пользователи таких систем могут быть незнакомы с языком, на котором написаны документы коллекции.

Для краткого представления содержания вышеперечисленных типов текстов требуются другие виды аннотаций. Такой аннотацией мог бы служить список наиболее частотных терминов текста. Но в таких списках термины, относящиеся к различным подтемам текста, перемешаны, что затрудняет их восприятие. Кроме того, большое количество терминов наиболее представительной темы может занять все предоставленное место, и термины других важных для текста тем будут упущены. Поэтому был предложен метод создания структурной тематической аннотации документа [108].

Структурная тематическая аннотация представляет содержание текста посредством описания участников его основной темы, которые моделируются совокупностью понятий, относящихся к этим темам. Структурная тематическая аннотация содержит наиболее информативные фрагменты тематического представления текста, которое включает все понятия текста, разбитые на тематические узлы.

Тематическое представление, содержащее все понятия текста, является слишком подробным, чтобы служить структурной аннотацией текста, предъявляемой пользователю. Поэтому в качестве структурной аннотации может быть использованы выделенные основные тематические узлы тематического представления.

В качестве примера рассмотрим (Рис. 4.12) структурную тематическую аннотацию Федерального закона об информации, информатизации и защите информации Российской Федерации (40 Кб, 164 различных термина).

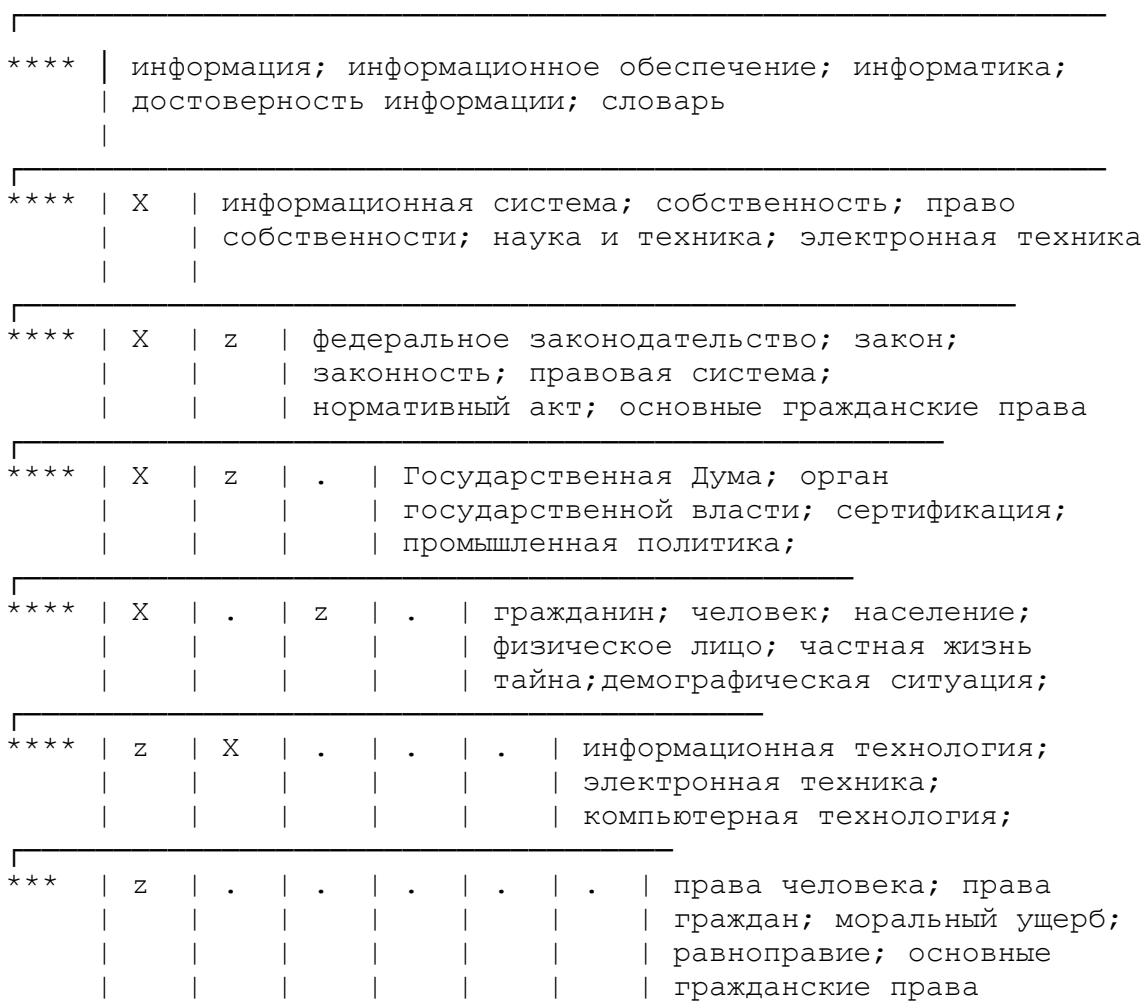


Рис. 4.12. Структурная аннотация для Федерального закона об информации, информатизации и защите информации Российской Федерации

Структурная тематическая аннотация включает в себя следующие части:

- понятия основных тематических узлов, упорядоченных в порядке убывания частотности и расположенных горизонтально;
- отметки об относительно суммированной частотности основных тематических узлов, обозначаемые различным количеством символов “*”;
- отметки об относительной силе взаимоотношений между различными тематическими узлами

– ”X“– очень сильное отношение;

– ”z”– сильное отношение;

– ”.” – отношение.

Структурная аннотация позволяет оценивать содержание текста с одного взгляда, в частности, служит хорошим инструментом для тестирования той конкретной лингвистической онтологии, на основе которой обрабатывается документ.

4.5.3. Построение аннотации для новостного кластера на основе тематического представления текстов кластера

Современные технологии обработки новостных потоков обычно включают в себя краткое представление содержания новостного кластера в виде аннотации (обзорного реферата) [292]. В данном разделе мы рассмотрим автоматический метод создания аннотации новостного кластера на основе тематического представления, построенного для этого кластера.

4.5.3.1. Построение тематического представления для новостного кластера

Новостной кластер представляет собой совокупность тематически близких документов. Поэтому тематическую структуру новостного кластера так же, как и отдельного элемента можно выявить за счет построения тематического представления этого кластера, и это представление можно будет использовать для управления набором предложений в аннотацию кластера, а именно для решения таких задач, как обеспечение полноты, снижения повторов, а также обеспечения связности аннотации кластера.

Построение тематического представления новостного кластера осуществляется простым способом: все тексты кластера склеиваются в единый текст, для которого производится стандартный тематический анализ одного документа и строится тематическое представление.

Результат этой процедуры, а затем и результат построения аннотации в некоторой степени зависит от порядка просмотра документов в кластере. Мы используем следующий метод объединения документов кластера в единый текст, используемый для построения аннотации.

Сначала в новостном кластере определяется «центр кластера» – документ, наиболее близкий к центру тяжести множества документов кластеров в метрическом пространстве нормализованных лемматическом и концептуальном (по понятиям лингвистической онтологии) индексов. Определяется «ядро» кластера – документы достаточно близкие к центру (по некоторому порогу). Затем «центр кластера» сдвигается в документ из ядра кластера, который был опубликован последним по времени. Пересчитываются веса связей документов кластера к новому центру. С учетом задаваемого интервала времени по убыванию веса сначала заполняются документы за последнее время, затем все остальные. Так как отбирается всего несколько предложений, то имеется общее ограничение на количество отбираемых в «единый документ» документов [292].

После порождения «единого документа» кластера для него строится тематическое представление. Так, для кластера, в который входит текст примера из раздела 4.5.1, основными тематическими узлами становятся следующие совокупности понятий (справа указана частотность понятия в кластере):

КИТАЙ	103
ПЕКИН	21
ГУАНЧЖОУ	13
ГОСУДАРСТВО	9
ЮАНЬ	7
ШАНХАЙ	6
КИТАЙЦЫ	5
НАНКИН	5
ГУАНДУН	1
ТАЙВАНЬ	103
ТАЙБЕЙ	21

АВИАЦИОННЫЕ ПЕРЕВОЗКИ	33
АВИАЦИОННАЯ КОМПАНИЯ	9
САМОЛЕТ	9
АВИАРЕЙС	7
ТРАНСПОРТНАЯ СФЕРА	4
АЭРОПОРТ	3
ТРАНСПОРТНЫЕ ПЕРЕВОЗКИ	2
АЭРОБУС	1
АВИАЛИНИЯ	1
ГОРОД	17
ТЕРРИТОРИЯ, УЧАСТОК	3
НАСЕЛЕННЫЙ ПУНКТ	1
ОСТРОВ	17
ЖИТЕЛЬ ОСТРОВА	1
ЧАРТЕРНЫЕ ПЕРЕВОЗКИ	14
ТУРИСТ	12
ЧЕЛОВЕК	62
ТУРИЗМ	2
ПОЕЗДКА	1
ПАССАЖИР	10
ПРАВИТЕЛЬСТВО	6
РУКОВОДИТЕЛЬ	6
ОРГАН ПУБЛИЧНОЙ ВЛАСТИ	3
РУКОВОДСТВО	2
ОРГАН ИСПОЛНИТЕЛЬНОЙ ВЛАСТИ	2
ПУБЛИЧНАЯ ВЛАСТЬ	1

Таким образом, по основным тематическим узлам тематического представления могут быть определены основные элементы, обсуждаемой в кластере темы.

Как видно, тематические узлы включают концепты достаточно разной частотности. Низкочастотные концепты тематического узла могут быть ошибочно включены в тематический узел, кроме того, представительность ими основной темы документа невелика. Поэтому можно задать выделение ядра тематических узлов, которое определяется как коэффициент от 0 до 1. Этот коэффициент определяет, какая доля наиболее частотных понятий от общей частотности понятий в тематическом узле будет включена в ядро.

Так, при значении коэффициента тематического ядра 0.7 получим следующие ядра тематических узлов:

КИТАЙ	103
ПЕКИН	21
ГУАНЧЖОУ	13
ТАЙВАНЬ	103
АВИАЦИОННЫЕ ПЕРЕВОЗКИ	33
АВИАЦИОННАЯ КОМПАНИЯ	9
САМОЛЕТ	9
ГОРОД	17
ОСТРОВ	17
ЧАРТЕРНЫЕ ПЕРЕВОЗКИ	14
ТУРИСТ	12
ЧЕЛОВЕК	62
ПАССАЖИР	10
ПРАВИТЕЛЬСТВО	6
РУКОВОДИТЕЛЬ	6
ОРГАН ПУБЛИЧНОЙ ВЛАСТИ	3

4.5.3.2. Метод построения аннотации новостного кластера по тематическому представлению кластера

Аннотация новостного кластера обычно состоит из заголовка и нескольких предложений из разных документов новостного кластера.

Зная ядра тематических узлов, полнота изложения содержания кластера обеспечивается тем, что нужно отбирать для аннотации предложения, содержащие пары этих тематических узлов – именно тогда эти предложения будут описывать взаимоотношения между основными тематическими элементами кластера.

При отборе заголовка для аннотации ищется заголовок, содержащий пару наиболее частотных тематических узлов. Если таких заголовков нет, то

ищутся заголовки, содержащие понятия из одного наиболее частотного тематического узла.

Для выбора очередного предложения в списке основных тематических узлов отмечаются все тематические узлы, которые уже были упомянуты. Очередное предложение должно содержать пару основных тематических узлов: наиболее частотный тематический узел, который еще не упоминался, и какой-нибудь еще основной тематический узел.

Для обеспечения связности требуется, чтобы очередное предложение содержало либо уже упомянутый тематический узел, либо уже упоминавшееся слово с большой буквы.

Кроме того, делается ряд дополнительных проверок:

- предложение не должно являться вопросительным или отрицательным предложением,
- предложение не должно содержать в заданном числе первых слов местоимение,
- начало предложения не должно совпадать с началами заголовка и предложений, уже взятых в аннотацию,
- число слов предложения, совпадающего со словами предшествующих предложений не должно превышать некоторой доли длины предложения.

Понятно, что даже при проверке вышеупомянутых условий может найтись еще достаточно много подходящих предложений-кандидатов. Кроме того, оценка предложений на основе понятий ЛО не является достаточной без учета упоминаемых именованных сущностей, которые могут быть и не описаны в базовой лингвистической онтологии.

Поэтому вводится еще и общая оценка предложения с помощью вычисления веса предложения, которая складывается из двух компонентов: весов упомянутых понятий ЛО, которые были получены в тематическом

представлении, а также весов содержащихся в предложении слов с большой буквы, не считая первого слова предложения.

Для вычисления весов слов с большой буквы (далее Слов), сначала вычисляется вес самого частотного Слова w_{max_word} в документе кластера:

$$w_{\max\_word} = \min(1, w_{\max\_conc} \cdot \frac{Fr_{\max\_word}}{Fr_{\max\_conc}})$$

где w_{max_conc} – максимальный вес понятия ЛО в тематическом представлении, Fr_{max_conc} – частотность в тексте понятия ЛО с максимальным весом, Fr_{max_word} – частотность самого частотного Слова.

Остальные веса Слов (w_{word}) вычисляются пропорционально их частотности:

$$w_{word} = w_{\max\_word} \cdot \frac{Fr_{word}}{Fr_{\max\_word}}$$

Так веса понятий и слов сводятся к одной шкале.

Просмотр предложений-кандидатов начинается с начала документа кластера, т.е. предложения набираются сначала из главного документа кластера и наиболее близких к нему по содержанию. Каждое следующее предложение берется из другого документа.

Для кластера примера была получена следующая аннотация (в скобках указан источник новости и время публикации):

Предложения	Тематические узлы
Китай и Тайвань установили авиасообщение после 60-летнего перерыва *(Новые Известия - лента новостей, 04.07.2008 11:08:45)*	*КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ (авиасообщение)*
Первый чартерный рейс с 250 пассажирами уже прибыл в столицу Тайваня из китайского города Гуанчжоу. *(Lenta.ru - главные новости, 04.07.2008 9:47:25)*	*КИТАЙ, ЧАРТЕРНЫЕ ПЕРЕВОЗКИ (чартерный рейс), ГОРОД, ПАССАЖИР*

Предложения	Тематические узлы
С 4 июля самолеты с материкового Китая на остров Тайвань и обратно будут летать каждую неделю с пятницы по понедельник. (РегKURSCITY.RU - Курс, 04.07.2008 9:35:34)	*КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ (самолет), ОСТРОВ*
Перед прибывающими в ближайшие выходные 600 туристами из Китая будет расстилаться красная ковровая дорожка. (BBCRussian.com (Главная), 04.07.2008 1:18:25)	*КИТАЙ, ТУРИСТ*
По завершении в 1949 году гражданской войны в Китае и изгнания правительства Гом-Инь-Дана на Тайвань, отношения между двумя сторонами Тайваньского пролива были заморожены. (РегЛІГАБізнесІнформ - Украины - Новости за рубежом, 04.07.2008 9:14:00)	*КИТАЙ, ТАЙВАНЬ, ПРАВИТЕЛЬСТВО*

В заголовке аннотации мы имеем три основных тематических узла: *КИТАЙ, ТАЙВАНЬ, АВИАЦИОННЫЕ ПЕРЕВОЗКИ:*

- в первом предложении сообщается о конкретных городах, связанных с авиаперевозками, и указывается о том, что перевозки чартерные – таким образом, упомянуты еще два тематических узла – *ГОРОД* и *ЧАРТЕРНЫЕ ПЕРЕВОЗКИ;*
- второе предложение содержит новый тематический узел *ОСТРОВ;*
- третье предложение содержит узел *ТУРИСТ;*
- четвертое предложение содержит тематический узел *ПРАВИТЕЛЬСТВО*

Таким образом, каждое предложение содержит не менее двух разных основных тематических узлов, один из которых новый (выделен подчеркиванием в правом столбце таблицы), а другой был упомянут ранее.

4.5.3.3. Тестирование предложенной модели аннотации новостного кластера

Предлагая метод аннотирования новостного кластера, мы сделали несколько предположений о внутренней структуре аннотации и о нашей способности выявлять эту структуру на основе создаваемого автоматически тематического представления. Для проверки предложенной модели аннотации новостного кластера был проведён эксперимент по проверке соответствия сделанных предположений ручным аннотациям, составленными экспертами-лингвистами.

Лингвисты создали несколько аннотаций новостных кластеров из предложений этого кластера. Аннотация представляла собой заголовок и четыре предложения. Общее количество разных аннотаций в эксперименте – 13. Для новостных кластеров были получены их тематические представления. Далее ручные аннотации были размечены на предмет наличия основных тематических узлов для данного кластера и именованных сущностей.

Задачей данной разметки являлась проверка описанных выше условий для составления аннотаций, а именно:

1. Действительно ли реальные аннотации должны содержать в себе как минимум два основных тематических узла из тематического представления текста и/или именованные сущности.
2. Используются ли в ручных аннотациях понятия-элементы основных тематических узлов и именованные сущности для организации лексической связности текста, а именно, повторяются ли в последующих предложениях ручных аннотаций понятия уже упомянутых основных тематических узлов или уже упомянутые именованные сущности.

3. Содержат ли очередные предложения элементы новизны в виде нового, еще не упоминавшегося тематического узла или именованной сущности.

Результаты эксперимента представлены в Табл. 4.4 [233].

Табл. 4.4. Выявление основных тематических узлов и именованных сущностей в ручных аннотациях

Проверка представленности основных тематических узлов:	
Всего предложений:	**65**
Из них количество предложений с не менее чем двумя тематическими узлами:	**60**
Количество предложений, в которых есть один основной тематический узел и не менее чем одна именованная сущность:	**58**
Оценка связности и новизны:	
Общее количество предложений, не считая первые предложения:	**52**
Количество предложений с новым основным тематическим узлом:	**35**
Количество предложений с новым именем:	**28**
Количество предложений с повтором упоминавшегося тематического узла:	**46**
Количество предложений с повтором упоминавшегося имени:	**38**

Результатом проведённого анализа явился тот факт, что 83% предложений реальных ручных аннотаций (от общего числа предложений), составленных экспертами-лингвистами, удовлетворяют сделанным

предположениям. Особенность оставшихся 17% предложений состоит в том, что все они являлись последними предложениями ручной аннотации. Такая ситуация связана с тем, что основная тема новостного кластера уже изложена, и дальнейшее описание событий «разрывается» на второстепенные темы документы, которых обычно имеется большое количество. Проведенный эксперимент доказывает, что сделанные предположения в методе автоматического аннотирования новостных кластеров имеют высокую корреляцию со структурой человеческих аннотаций.

4.5.3.4. Оценка качества аннотаций новостных кластеров

Как уже упоминалось в разделе 4.3.2, тестирование качества автоматических аннотаций является сложной процедурой. В качестве метрики аннотаций новостных кластеров, позволяющая автоматизировать этот процесс, используется такая метрика, как ROUGE, которая подсчитывает число перекрытия (n-граммы слов) автоматической аннотации с «идеальными» аннотациями, составленными людьми [104].

Другой используемой мерой оценки качества аннотаций является Метод Пирамид, который основан на ручном выделении экспертами «информационных единиц» из эталонных аннотаций – Summary Content Units (SCUs) и вычислении процентной доли этих единиц, упомянутых в автоматических аннотациях [72].

Далее рассмотрим подробнее результаты применения этих методов оценки для тестирования наших аннотаций новостных кластеров. Кроме того, будет рассмотрена процедура применения ручных оценок.

4.5.3.4.1. Тестирование аннотаций новостных кластеров методом ROUGE

Поскольку в разных статьях, описывающих эту метрику, содержатся несколько разные способы ее вычисления, то конкретные используемые

формулы были названы ROUGE-1-cir и ROUGE-2-cir [292] и вычислялись следующим образом:

$$ROUGE-N-cir(A_i) = \frac{\sum_{M_{ij}} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_{ij}} count(Ngram(M_{ij})} ,$$

где A_i – оцениваемая обзорная аннотация *i*-того кластера, M_{ij} – ручные аннотации *i*-того кластера, *Ngram(D)* – множество всех n-грамм из лемм соответствующего документа *D*. При сравнении отдельных документов в расчет берутся только уникальные n-граммы, присутствующие в обоих документах – не поощряется многократный повтор одного и того же предложения. При рассмотрении нескольких аннотаций, наоборот, повторение одинаковых элементов поощряется. Биграммы учитывались с перестановками.

Для оценки качества построенных аннотаций были использованы данные, предоставленными С.Д. Тарасовым (Военмех, Спб.). В проведенных С.Д. Тарасовым экспериментах группе студентов было предложено построить ручную аннотацию для новостных кластеров, которые брались из системы Google.Новости в период с 01 по 05 декабря 2008 года. Ручная аннотация должна была быть составлена из четырех предложений. Ограничений на выбор предложений из разных текстов не накладывалось.

Случайным образом были выбраны 15 новостных кластеров разной тематики, включая новости о погоде, спорте, финансах и политике, для которых имелось от 18 до 40 ручных аннотаций (всего 462). В качестве «базовой оценки», следуя [41], рассматривались следующие варианты искусственных аннотаций: первый документ кластера; заголовки первых четырех документов; первые предложения первых четырех документов; последний документ кластера.

В качестве автоматической аннотации рассматривались аннотации из заголовка и трех предложений, взятых из разных текстов. Были получены

следующие результаты (в Табл. 4.5 приведены результаты для разных параметров ядра кластера – см. п. 4.5.3.2):

Вид аннотации	ROUGE-1-cir	ROUGE-2-cir
первый документ кластера	0.219	0.083
заголовки первых четырех документов	0.162	0.056
первые предложения первых 4 документов	0.269	0.107
последний документ кластера	0.278	**0.168**
автоматическая аннотация с ядром 0,20	**0.331**	0.150
автоматическая аннотация с ядром 0,40	0.328	0.140

Табл. 4.5. Результаты тестирования различных видов аннотаций новостных кластеров по методу ROUGE.

Следует отметить, что некоторые ручные аннотации совпадали с первым или последним документом кластера. Определенным недостатком используемых данных является то, что некоторые кластеры содержали документы за несколько дней, поэтому ручные аннотации чаще содержали предложения из последних документов кластера.

Существует определенная критика использования метрик ROUGE для оценки качества аннотирования. Метрика чувствительна к длинам сравниваемых документов, не учитывает связность аннотаций. В целом, существует большое разнообразие между ручными аннотациями разных экспертов. В нашем случае нам лишь важно было оценить близость построенных автоматических и ручных аннотаций для оценки перспективности предложенного подхода.

4.5.3.4.2. Тестирование аннотаций новостных кластеров методом Пирамид

Метод Пирамид основан на выделении в аннотациях отдельных единиц получаемой информации (SCU) [72]. Выделенная информационная единиц» получает вес, равный количеству ручных эталонных аннотаций, где она встречается. Название «Метод пирамид» как раз и связано с тем, что информационные единицы SCU выстраиваются как бы в пирамиду: на вершине небольшое число единиц с большим весом, внизу пирамиды – большое число информационных единиц с маленьким весом. Общая оценка автоматической аннотации складывается из суммы весов SCU, которые она содержит, по отношению к общему количеству SCU для данного текста:

$$\frac{[Суммарный\_вес\_найденных\_SCU]}{[Суммарный\_вес\_всех\_SCU\_для\_данного\_топика]}$$

В качестве примера SCU и её вхождений в тексты аннотаций можно рассмотреть следующие фрагменты предложений новостного кластера:

SCU: Мини-субмарина попала в ловушку под водой.

1. *мини-субмарина... была затоплена... на дне моря...*
2. *маленькая... субмарина... затоплена... на глубине 625 футов.*
3. *мини-субмарина попала в ловушку... ниже уровня моря.*
4. *маленькая... субмарина... затоплена... на дне морском...*

Для сравнения качества предложенного метода аннотирования новостных кластеров в терминах информационных был реализован известный метод аннотирования Maximal Marginal Relevance (MMR) [28], показавший высокое качество аннотирования на конференции SUMMAC, а его модификации и на более поздних конференциях. Метод MMR – это итеративный алгоритм выбора предложений в аннотацию. Пусть имеются:

- Q – запрос для аннотирования или в нашем случае общего тематического аннотирования – вектор слов всего кластера,
- S – множество предложений кандидатов,
- s – рассматриваемое предложение кандидат,
- E – множество выбранных предложений.

Тогда на каждой итерации предложение в итоговую аннотацию будет отбираться в соответствии с формулой:

$$MMR = \arg\max_{s \in S} \left[\lambda \cdot Sim_1(s, Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) \right]$$

Предложения итоговой аннотации сортируются в соответствии с их порядком следования в исходном документе.

Для предложенного нами метода аннотирования новостного кластера и метода MMR была применена пирамидная оценка. Сравнивались аннотации длиной 100 слов. Предложенный метод аннотирования получил среднюю оценку 63.8%, метод MMR – 64.3%. Таким образом, по полноте изложения информации предложенный метод не показал лучшие результаты. На наш взгляд, это частично связано с тем, что для обеспечения лучшей связности аннотации требуется некоторая степень повторяемости в предложениях.

4.5.3.4.3. Оценка связности аннотаций новостных кластеров

Тестирование связности и читабельности автоматических аннотаций может производиться только человеком. Была применена следующая процедура: лингвист должна была читать каждый вид аннотации последовательно от предложения к предложению, и каждому предложению выставить некоторый штрафной балл:

0.0 – если предложение «хорошее» (связано с остальными предложениями, качественно вписывается в данную аннотацию и т.д.),

1.0 – если предложение «плохое» (не связано с другими предложениями, является лишним в данной аннотации и т.д.)

0.5 – в спорных ситуациях.

Таким образом, каждый вид аннотации получил некоторую совокупность штрафных баллов, чем меньше баллов, тем лучше. В среднем аннотации, порожденные методом MMR, получили 0.7 штрафных баллов, предложенным методом – 0.3 балла [233].

Таким образом, описанные методы аннотирования отдельного документа и новостного кластера на основе тематического представления позволяют решать такие проблемы методов автоматического аннотирования как обеспечение полноты представления содержания, снижения повторов, обеспечения связности аннотации. Основная суть предложенных методов автоматического аннотирования заключается в выявлении основных участников обсуждаемой в тексте или кластере ситуации и в предположении, что наиболее информативными являются предложения, в которых сообщается информация о взаимодействии этих сущностей.

Полнота передачи содержания документа (документов) обеспечивается тем, что отбираются предложения, упоминающие основных участников ситуации. Снижение повторов становится возможным, поскольку один и тот же участник ситуации может быть распознан в значительном разнообразии текстовых выражений. Кроме того, снижение повторов обеспечивается обязательным упоминанием нового, еще не упомянутого элемента тематического представления в очередном предложении аннотации. Наконец, связность аннотации обеспечивается повторяемостью тематических узлов и именованных сущностей.

Выявленные закономерности построения аннотаций новостных кластеров не обязательно требуют наличие лингвистической онтологии в основе обработки. Нахождение основных участников ситуации может быть смоделировано на основе совершенно других нетезаурусных методов обработки текстов, а фактор необходимости упоминания в предложениях аннотации, по крайней мере, двух основных участников может быть

добавлен как фактор в совокупность учитываемых факторов, таких, как вес предложения, сходство с заголовком, позиционное расположение и др.

4.6 Применение предложенных методов для автоматической обработки текстов в различных проектах

4.6.1.Программный комплекс АЛОТ

Описанные лингвистические ресурсы и методы обработки текстов объединены в программный комплекс АЛОТ (Автоматизированная Лингвистическая Обработка Текстов). Программно-лингвистический комплекс АЛОТ производит автоматическую обработку поступающих на вход информационной системы потоков документов. Получая на входе файлы в формате HTML, АЛОТ на выходе выдает текстовые файлы в специальном формате, содержащие морфологический (нормализованные слова документа) и тематический индексы (термины и рубрики), предназначенные для дальнейшей загрузки в базу данных [226].

АЛОТ включает следующие этапы автоматизированной лингвистической обработки текстов:

- Морфологический анализ, в ходе которого анализа всем словам анализируемого текста сопоставляется грамматическая информация (род, число, падеж, категория одушевленности и т.п.).
- Автоматический тематический анализ, включая разрешение лексической многозначности, построение тематического представления текста, формирование концептуального индекса понятий ЛО с весами, полученными на основе построенного тематического представления (рис. 4.13),

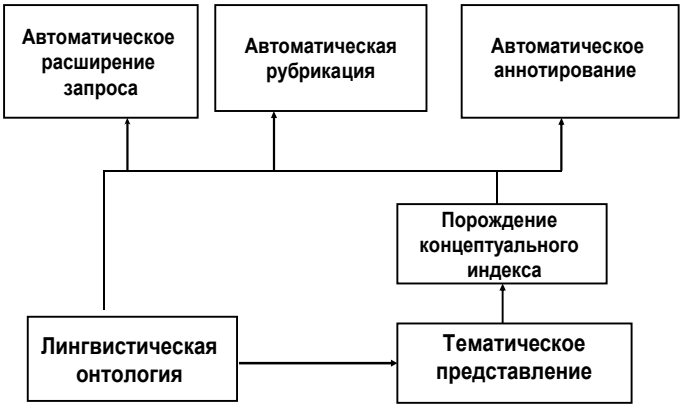


Рис. 4.13. Схема приложений автоматической обработки текстов на основе ЛО

- Автоматическое рубрицирование,
- Автоматическое аннотирование.

Программно-лингвистический комплекс АЛОТ использовался в более чем двадцати проектах по автоматической рубрикации текстов, включая, в частности, такие системы рубрикации, как:

- рубрикация законодательных актов по Классификатору правовых актов РФ – 1169 рубрик,
- рубрикация научных статей по экономике по рубрикатору JEL (ссылка – 700 рубрик),
- рубрикация по правовому классификатору Центральной избирательной комиссии (450 рубрик, 4 уровня),
- рубрикация социологических опросов по рубрикатору 300 рубрик и др.

Программно-лингвистический комплекс АЛОТ используется при комплексной обработке потоков новостей Интернет-сервиса Рамблер, включая морфологический анализ новостных сообщений, формирование

концептуального индекса, автоматическую рубрикацию по нескольким рубрикаторам, аннотирование отдельного документа и новостного кластера.

В проектах с Банком России комплекс АЛОТ используется для рубрикации потоков новостей по двум банковско-финансовым рубрикаторам и порождения аннотаций отдельных документов и новостных кластеров.

Наиболее комплексно АЛОТ используется для обработки документов и визуализации выдачи в Университетской информационной системе Россия (www.cir.ru).

4.6.2. АЛОТ в УИС РОССИЯ

Университетская информационная система Россия (УИС РОССИЯ) (www.cir.ru) создана и развивается как тематическая электронная библиотека и база для исследований и учебных курсов в области экономики, управления, социологии, лингвистики, философии, филологии, международных отношений и других гуманитарных наук [236]. Развитие УИС Россия было начато в 1993 году под руководством Юдиной Т.Н., Журавлева С.В., Леонтьевой Н.Н. [226].

УИС Россия содержит более 2 млн. документов, включая нормативно-правовые акты, издания государственных органов, публикации СМИ. Каждый поступающий в систему документ обрабатывается программным комплексом АЛОТ (рис. 4.14). Разработанный на основе описанной модели Общественно-политический тезаурус служит как поисковое средство и средство визуализации поисковой выдачи в УИС РОССИЯ.

В рамках УИС РОССИЯ АЛОТ производит автоматическую рубрикацию по двум рубрикаторам:

- по рубрикатору нормативных актов, разработанному в Центре информационных исследований (180 рубрик, 3 уровня иерархии);
- по рубрикатору Исследовательской службы конгресса Библиотеки конгресса США (Legislative Indexing Vocabulary, LIV (80 рубрик);

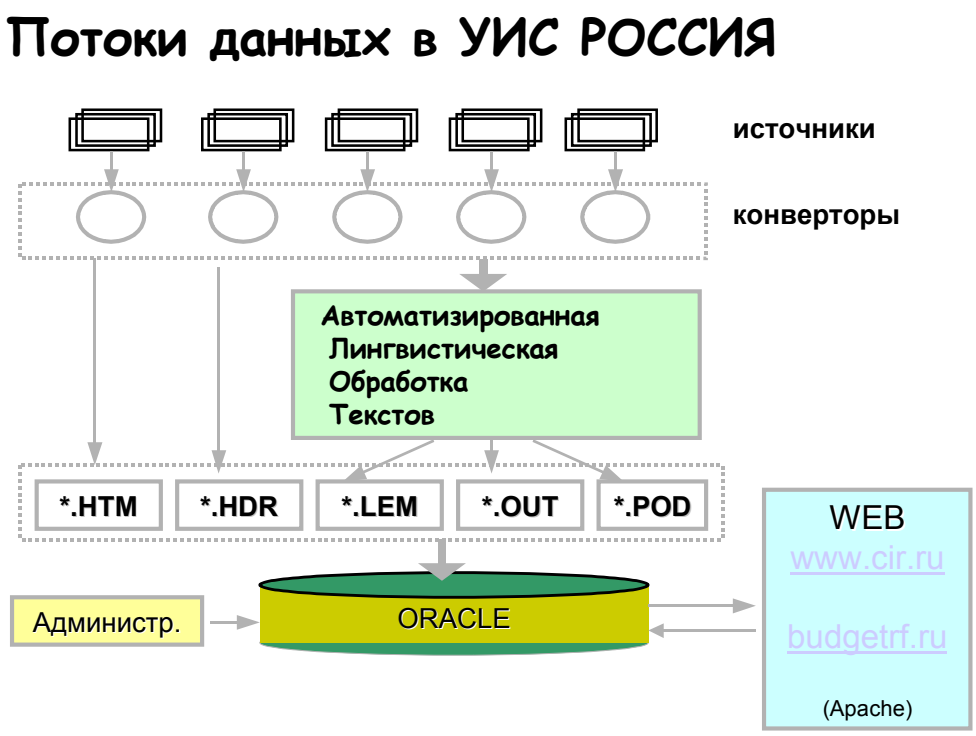


Рис. 4.14. Схема обработки документов в УИС РОССИЯ

4.6.3. Общественно-политический тезаурус как поисковое средство в УИС РОССИЯ

В качестве базовой лингвистической онтологии для УИС РОССИЯ служит Общественно-политической тезаурус [250].

Для использования в поиске по документам УИС Россия знаний из Общественно-политического тезауруса пользователь может задать булевский запрос, включающий как слова, так и понятия тезауруса. Понятие тезауруса может быть задано без расширения по дереву. Тогда в ответ на запрос будут выданы документы, содержащие хотя бы одно из текстовых выражений, сопоставленных данному понятию.

Если понятие тезауруса задано с расширением по дереву, то релевантными считаются документы, содержащие хотя бы один синоним выбранного понятия или (с несколько меньшим весом) хотя бы один синоним понятий из дерева-вниз выбранного понятия. Таким образом, выбор

в запрос одного понятия может оказаться равносилен выбору сотен и тысяч слов и словосочетаний.

Использование опции “расширение по дереву Тезауруса” при поиске с использованием географических названий позволяет найти все географические названия и административные единицы. При поиске по термину *ЮГО-ВОСТОЧНАЯ СИБИРЬ* будут выданы также документы, содержащие: *БАЙКАЛ, ЗАБАЙКАЛЬЕ, БУРЯТИЯ, ЧИТИНСКАЯ ОБЛАСТЬ, ПРИБАЙКАЛЬЕ* и т.д.

Особенно впечатляющих результатов удается добиваться, формируя запрос из нескольких понятий с расширением по дереву. В частности, можно эффективно найти документы следующей тематики:

```
/Термин_расш=“ПРЕСТУПНОСТЬ”
and /Термин_расш= “СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ ОКРУГ”
```

или, например,

```
/Термин_расш=“МИГРАЦИЯ”
and /Термин_расш= “АМУРСКАЯ ОБЛАСТЬ”
```

Общественно-политический тезаурус играет большую роль для визуализации поисковой выдачи в УИС Россия. При формировании выдачи документов на запрос, происходит выявление наиболее характерных для данной выдачи понятий тезауруса. Выделенные понятия выдаются на экран в виде так называемого информера. Понятия тезауруса в информере упорядочиваются на основе веса, полученного по формуле типа tf.idf, когда частотность упоминания понятия в выдаче сопоставляется с частотностью упоминания понятия в коллекции. Таким образом, информер позволяет пользователю увидеть внутренние темы текущей поисковой выдачи и в случае необходимости модифицировать запрос.

Информационные системы типа УИС Россия с соответствующими предметно-ориентированными лингвистическими онтологиями и средствами

загрузки своих документов поставляются в рамках проектов с другими организациями.

Заключение к главе 4

В данной главе были представлены методы и алгоритмы применения лингвистических онтологий для автоматической обработки текстов в различных приложениях информационного поиска, включая новые методы

- автоматического разрешения лексической многозначности на основе локального и глобального контекстов,

- автоматического построения булевского запроса для вопроса типа «формулировка проблемы»,

- автоматической рубрикации, которая основывается на структуре лингвистической онтологии и тематическом представлении текста,

- автоматического аннотирования одного и многих документов, на основе тематического представления текстов.

Представлены эксперименты, показывающие важный вклад дополнительной языковой и онтологической информации, по сравнению с пословными методами решения задач.

Глава 5. Многофакторная модель автоматического извлечения терминов предметной области

Важным аспектом создания прикладных систем в конкретных предметных областях является учет терминологии предметной области. Поэтому одним из важных направлений исследований в области создания понятийных моделей предметных областей, прикладных онтологий являются технологии извлечения из текстов терминов предметной области. Термины предметной области соответствуют понятиям этой области [48, 241, 242, 247, 307, 314, 315] и являются также необходимой базой для создания предметно-ориентированных баз знаний.

Как известно, под термином понимается слово (или сочетание слов), являющееся точным обозначением определенного понятия какой-либо специальной области науки, техники, искусства, общественной жизни и т.п. [271]. Понятно, что такое определение невозможно применить для автоматического извлечения терминов из текстов, поэтому на практике для отбора терминов применяется некоторый набор лингвистических и статистических характеристик словосочетаний.

5.1. Необходимость разработки многофакторной модели для извлечения терминов

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

- статистические характеристики употребления словосочетания и его компонентов (частотность по коллекции, взаимная ассоциация, вхождение в объемлющие словосочетания и т.п.);

- синтаксические ограничения: извлекаются словосочетания заданной синтаксической структуры: группы прилагательное + согласованное

существительное, существительное + существительное в родительном падеже и др.;

- лексические фильтры, например, не извлекаются словосочетания, включающие географические названия, эмоциональную лексику и др. [272].

Методы извлечения терминов значимым образом варьируются в зависимости от длины термина, часто отдельно рассматриваются:

- извлечение однословных терминов,

- извлечение двухсловных терминов,

- извлечение трех и более многословных терминов.

Особенности извлечения разных типов терминов связаны с частотностью встречаемости того или иного типа терминов, сложностью их синтаксического отграничения в тексте и др.

Так, анализ имеющихся терминологических словарей и тезаурусов показывает, что основную массу тезаурусных единиц составляют слова-существительные, а также словосочетания из двух-трех слов. Наиболее часто структура словосочетаний основывается на зависимых от главного существительного прилагательных и существительных в родительном падеже. В то же время, например, доля терминов с предлогами относительно невелика, при этом имеется большая проблема неоднозначности в выделении предложных конструкций из текста.

Работа с терминологиями реальных предметных областей показывает, что эксперты при отборе терминов в словари, информационно-поисковые тезаурусы руководствуются набором критериев. Такие критерии особенно подробно разрабатывались в стандартах на разработку и ведение информационно-поисковых тезаурусов.

Поскольку в текстах предметной области может встречаться достаточно много частотных словосочетаний, то обычно стандарты на тезаурусы вводят правила включения терминологических словосочетаний в тезаурусы. Так, ГОСТ 7.25 [245] указывает, что допускается включать

словосочетания в словник тезауруса, если в качестве опорного слова они содержат существительное, и если выполнено одно из следующих условий:

- значение словосочетания не выводится из значений его компонентов, например,
  черный ящик, абсолютно черное тело, царская водка;
- хотя бы один из компонентов словосочетания не употребляется в составе других сочетаний или употребляется всегда в другом смысле, например, *торговля на вынос, легкая промышленность*;
- для данного словосочетания в словнике тезауруса существуют полные синонимы, например, *натрия хлорид = поваренная соль*;
- данное словосочетание является устойчивым словосочетанием с именем собственным: *таблица Менделеева, закон Бойля-Мариотта*;
- отдельные слова словосочетания имеют слишком широкое значение, например, слово *машины* в словосочетаниях: *строительные машины, электрические машины;*
- для данного словосочетания в словнике тезауруса существует общепринятая аббревиатура, например: *поверхностно-активные вещества – ПАВ, Универсальная десятичная классификация – УДК, информационно-поисковый тезаурус – ИПТ, электронно-вычислительная машина - ЭВМ;*
- разбиение словосочетаний на отдельные компоненты приводит к потере важных для поиска семантических связей. Так, разбиение языкового выражения *язык программирования* не позволяет установить связи с такими языковыми выражениями как *Алгол, Кобол, Фортран*.

Словосочетания, которые не удовлетворяют перечисленным условиям, рекоменудется разбивать на компоненты.

Американский стандарт [223] помимо вышеперечисленных случаев приводит также критерий общепринятости термина профессиональным

сообществом, например, *data processing – обработка данных.* Кроме того, этот стандарт указывает, что введение многословного дескриптора позволяет избегать ложных корреляций, например, разбиение термина *library science* (наука о библиотеках = библиотековедение), может привести к нахождению нерелевантных документов о научных библиотеках (*science library*).

Похожие принципы описываются в работе [316] на примере терминов научно-технического тезауруса [317].

Помимо вышеуказанных принципов отбора терминов, из определения следует, что термин должен быть достаточно частотен в текстах предметной области, и относительно мало частотен в текстах других предметных областей. Соответствие термина важному понятию предметной области может выражаться различными способами: частотностью, упоминанием в заголовках (например, интернет-страниц), толкованиями в известных терминологических словарях и др. Кроме того, если происходит процесс пополнения уже существующего словаря (тезауруса), то отдельный набор факторов может базироваться на особенностях терминов, уже включенных в словари (тезаурусы).

Таким образом, и автоматический отбор терминов должен базироваться на многофакторных моделях. Эффективность подхода для извлечения устойчивых словосочетаний на основе комбинирования признаков была продемонстрирована в ряде работ [50, 160, 206]. Вместе с тем, такие многофакторные модели должны быть переносимы с одной предметной области на другую. В данном исследовании предлагается подход по выявлению большого количества признаков для автоматического извлечения терминов из текстов и комбинирования этих признаков методами машинного обучения.

5.2. Особенности многофакторной модели извлечения терминов

5.2.1. Основные типы признаков для извлечения терминов

В предлагаемой модели используется три типа признаков для извлечения терминов [44, 116]:

- признаки, построенные на основе текстовой коллекции предметной области;

- признаки, полученные на основе информации глобальной поисковой машины,

- признаки, полученные на основе заданного тезауруса предметной области. Здесь моделируется ситуация развития существующего тезауруса и хотим выяснить, насколько знания, описанные в текущей версии тезауруса, могут улучшить качество автоматического извлечения следующих терминов.

Наборы признаков отличаются для отдельных слов, словосочетаний из двух слов и словосочетаний с большим количеством слов.

Обращение за дополнительной информацией к глобальным поисковым машинам важно по нескольким причинам.

Во-первых, коллекции документов широкой предметной области всегда недостаточно, поскольку множество достаточно значимых терминов предметной области может иметь относительно низкую частотность в данной коллекции. Привлечение Интернета помогает получить дополнительную информацию по таким словосочетаниям.

Во-вторых, использование информации из Интернета позволяет выяснить, насколько употребление данного слова или словосочетания жестко связано с заданной предметной областью.

Наконец, обращение в Интернет – это достаточно простой путь получения контекстов употребления слов исловосочетания. В качестве таких контекстов используются сниппеты (аннотации документов в выдаче), получаемые от поисковой машины Яндекс через xml-интерфейс. Так по

запросу *внешний долг* сниппетом одного из выданных документов является следующий:

Вне́шний долг *— суммарные денежные обязательства страны, выражаемые денежной суммой, подлежащей возврату* ***внешним*** *кредиторам на определенную дату, то есть общая задолженность страны по* ***внешним*** *займам и невыплаченным по ним процентам.*

Третий тип признаков – признаки, полученные на основе ЛО, представляют собой необычный тип признаков, однако использование этих признаков является чрезвычайно важным. Обычно в начале работы с терминологией предметной области несколько десятков наиболее существенных терминов являются очевидными, приводятся во всех терминологических словарях предметной области, часто достаточно понятно, как описать взаимоотношения между ними в терминологическом ресурсе. Чем больше производится работа над терминологически ресурсом широкой предметной области, тем сложнее его пополнять. Поэтому представляется важным, чтобы термины, которые уже отобраны экспертами в тезаурус, онтологию, терминологический словарь, помогали выявлять остальные термины данной предметной области.

Для комбинирования выделенных признаков для наилучшего извлечения терминов предметной области применяются методы машинного обучения.

5.2.2. Математические методы для комбинирования факторов

Задачей применения методов является переупорядочение исходного списка слов (первоначально упорядоченного по мере снижения частотности) так, чтобы в начало списка попало как можно больше слов-терминов. Таким образом, наилучшее переупорядочение списка снизит трудозатраты эксперта по вводу терминов в терминологические ресурсы – эксперт будет меньше просматривать слова, не являющиеся терминами. Для оценки качества такого

упорядочения используется мера, заимствованная из информационного поиска – так называемая средняя точность – AvP [227].

Характеристика средней точности **AvP** в задаче извлечения слов-терминов вычисляется следующим образом. Пусть в упорядоченном списке слов имеется k терминов, и $pos(i)$ – позиция ***i***-го термина от начала списка. Тогда точность на уровне ***i***-го термина $PrecTerm_i$ в упорядоченном списке равна $PrecTerm(pos(i))$, т. е. величина точности $PrecTerm_i$ подсчитывается в момент поступления в список i-го термина и равна доле терминов в списке от 1 до $pos(i)$ позиции**.** Средняя точность для данного упорядочения списка слов равна среднему значению величины $\boldsymbol{PrecTerm_i}$:

$$\boldsymbol{AvP} = \frac{1}{k}\sum_{1}^{k} \Pr ecTerm_i$$

Данная мера позволяет оценить качество извлечения терминов с помощью одной числовой величины за счет того, что, чем большая доля терминов из списка сосредоточена в начале списка, тем эта мера выше.

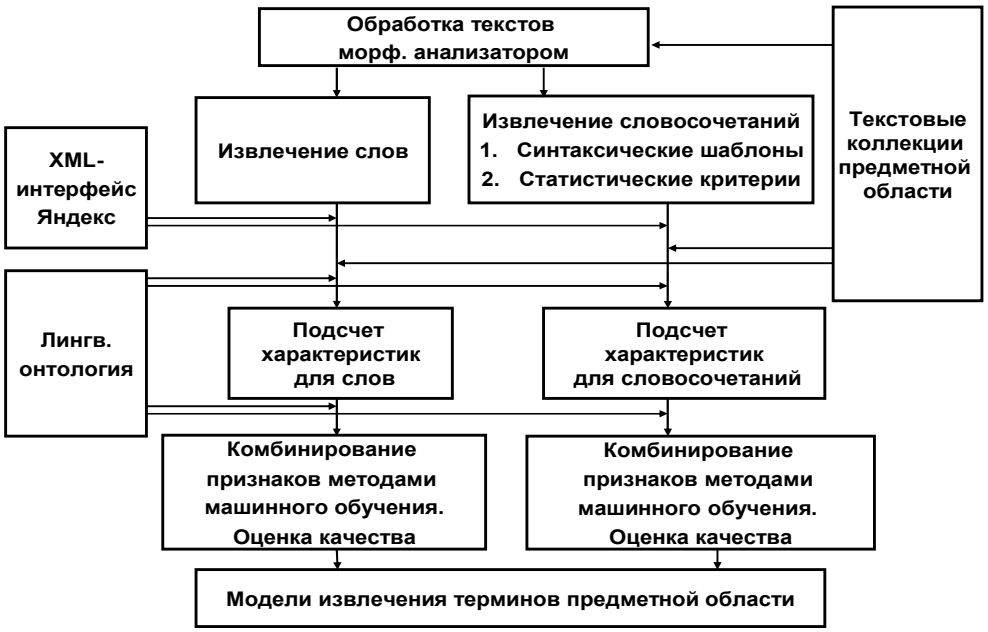


Рис. 5.1. Схема предложенного метода автоматического извлечения терминов. Лингвистические онтологии используются на двух этапах: как один из источников признаков и как средство для оценки качества извлечения

Рис. 5.1 представляет схему предложенного метода автоматического извлечения терминов. Для нахождения комбинации признаков, наилучшим образом отделяющих термины от нетерминов, используется метод машинного обучения – логистическая регрессия.

5.2.3. Логистическая регрессия как метод машинного обучения

Логистическая регрессия представляет собой метод построения линейного классификатора, позволяющий оценивать апостериорные вероятности принадлежности объектов классам некоторой системы классификации [232, 238].

Для применения метода для извлечения терминов предполагается, что языковые объекты (слова и выражения) описываются n числовыми признаками f_j: $X \to R$, $j=1,.., n$. Тогда пространство признаковых описаний объектов есть $X = R^n$.

Пусть Y – множество меток терминологичности {термин, нетермин} системы извлечения треминов, и задана обучающая выборка пар «языковое выражение, метка терминологичности» $X^m=\{(x_1, y_1).. (x_m, y_m)\}$. Y имеет два значения $\{-1, +1\}$: *«+1»* соответствует термину, *«-1»* – нетермину.

В методе логистической регрессии строится линейный алгоритм классификации α: $X \to Y$, вида

$$\alpha(x, w)=sign\ (\sum_{j=1}^{n} w_i f_i\ (x) - w_o\ ) = <x, w>$$

где w_j – вес j-го признака, w_0 – порог принятия решения, w = (w_0, $w_1,..w_n$) – вектор весов, $<x, w>$ – скалярное произведение признакового описания объекта на вектор весов. Предполагается, что искусственно введен константный нулевой признак: $f_0\ (x)=-1$.

Для построения классификатора решается задача минимизации эмпирического риска с функцией потерь вида:

$$Q\ (w) = \sum_{1}^{m} \ln(1+\exp(-y_i < x_i, w >)) \to \min_w$$

После нахождения *w* становится возможным находить класс объекта *x:*

$\alpha(x) = sign \langle x, w \rangle$

Кроме того, можно оценивать апостериорные вероятности принадлежности объекта *x* классам следующим образом:

$P(y|x) = \sigma(-y \langle x, w \rangle), y \in Y$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ – сигмоидная функция.

На практике данная апостериорная вероятность часто трактуется как оценка удаленности объектов от границ классов. Такая оценка позволяет, в нашем случае, упорядочивать языковые выражения в порядке снижения вероятности отнесения этого выражения к классу терминов и применять меру оценки качества извлечения терминов – *AvP*.

Подбор коэффициентом логистической регрессии может осуществляться посредством численных методов: градиентный спуск, метод Ньютона, стохастический градиентный спуск. В данной работе использовалась реализация метода логистической регрессии из пакета RapidMiner (http://rapid-i.com).

5.3. Постановка эксперимента по оценке качества извлечения словосочетаний. Используемые терминологические ресурсы

Для оценки качества предложенного комбинированного метода извлечения терминов эксперименты проводились в двух предметных областях, а результаты работы метода сравнивались с двумя вручную созданными терминологическими ресурсами. Одной из областей является широкая область по естественным наукам и технологиям, второй областью – банковская область. Для каждой предметной области имеются соответствующие текстовые коллекции, из которых извлечены слова и словосочетания – кандидаты в термины.

В области естественных наук используется онтологию OENT (Онтология по естественным наукам и технологиям) [261]. Эта онтология

представляет собой так называемую лингвистическую онтологию, т.е. онтологию, понятия в которой основаны на значениях существующих языковых выражений – в данном случае естественнонаучных терминов. Одновременно онтология OEHT может рассматриваться как тезаурус, поскольку описывает формализованные отношения между терминами предметной области. В настоящее время онтология OEHT включает 56 тысяч понятий и 150 тысяч различных терминов математики, химии, физики, геологии, биологии.

В банковской области используется тезаурус банковской деятельности, созданный по контракту с Центральным банком Российской Федерации. Он охватывает такие вопросы, как банковская деятельность, банковский контроль, кредитно-денежная политика, макроэкономика. В настоящее время тезаурус включает около 15 тысяч терминов.

Оба терминологических ресурса разработаны по модели лингвистической онтологии, предназначенной для автоматической обработки текстов в приложениях информационного поиска, описанной в главе 2.

Далее будут рассмотрены предлагаемые методы для извлечения однословных терминов и двухсловных терминов.

5.4. Метод отбора однословных терминов

При извлечении терминов предметной области большое внимание уделяется извлечению терминологических словосочетаний, и значительно меньшее исследований посвящено извлечению отдельных слов-терминов [225]. Вместе с тем известно, что список самых частотных словосочетаний, извлеченных из текстов предметной области, содержит очень высокую долю терминологических словосочетаний. В то время как подавляющее число наиболее частотных слов, извлеченных из коллекции текстов предметной области, представляют собой слова литературного языка, и, следовательно, не являются терминами. Применение статистических мер, позволяющих оценить особенность употребления слов в данной коллекции по сравнению с

некоторой контрастной коллекцией документов таких, как tf.idf или странность (weirdness) [225] повышает долю слов-терминов, получивших высокие веса по этим мерам, однако все еще остается относительно низкой.

Другой мерой, которая может применяться для выделения терминологических слов, является мера, оценивающая их вхождение как фрагмента в объемлющие словосочетания [10].

Рассмотрим признаки, которые можно использовать для выявления терминологичности слова, встретившегося в коллекции текстов предметной области.

5.4.1. Признаки, полученные на коллекции текстов предметной области

Частотность (Freq). Частотность употребления слова в коллекции.

Частотность с учетом частоты употребления в объемлющей коллекции (Tf.Idf). Данный признак широко употребляется в информационно-поисковых системах и позволяет снижать вес употребительных слов.

$$Tf.idf\,(w) = Tf\,log\,((n\text{-}b)/b)$$

где Tf – это частота употребления слова в текущей коллекции, n – размер контрастной коллекции, b – число документов, в которых употреблялось слово w в контрастной коллекции. В качестве контрастной коллекции для данного признака была выбрана коллекция Интернет-страниц белорусского Интернета, которая распространяется в качестве базовой коллекции для экспериментов в Интернет-поиске в рамках семинара РОМИП [3010].

Признак Странность (Weirdness). Данный признак учитывает пропорциональное соотношение частотности употребления слова в рабочей текстовой коллекции по сравнению с контрастной коллекцией [8]. Пусть w – слово. Тогда

$$Weirdness\ (w) = \frac{\frac{W_s}{T_s}}{\frac{W_g}{T_g}}$$

где W_s – частотность слова в коллекции предметной области; T_s – совокупная частотность слов в коллекции предметной области; W_g – частотность слова в контрастной коллекции белорусского интернета; T_g – совокупная частотность слов в контрастной коллекции белорусского интернета.

Признак C-Value. Данный признак основывает рейтинг терминологичности слов с учетом частотности объемлющих словосочетаний, в которое входит данное слово [10]. Пусть w – слово. Тогда

$$C\text{-}Value(w) = freq(w) - \frac{\sum_{mw \in T_a} freq\ (mw)}{|\ T_a\ |},$$

где T_a – множество всех словосочетаний *mw* в коллекции, содержащих слово w; $|T_a|$ – мощность множества T_a.

Наиболее частотное объемлющее словосочетание (Inside). Данный признак учитывает частотность наиболее частотного словосочетания, в состав которого входит данное слово. Пусть w – слово. Среди всех словосочетаний *mw*, содержащих слово w, выберем наиболее частотное. Тогда

$$Inside\ (w) = \frac{\max_{mw}\ (freq\ (mw))}{freq\ (w)}$$

Данный признак проверяет, не употребляется ли данное слово в составе одного и того же словосочетания. Чем выше значение признака, тем ниже вероятность того, что слово является самостоятельным значимым элементом предметной области, а, скорее, является компонентом более длинного устойчивого словосочетания.

Признаки употребления слова в наборе словосочетаний (Sum3, Sum10, Sum50). Данные признаки проверяют, насколько данное слово было

продуктивным в образовании словосочетаний предметной области. Пусть w – слово. Среди всех словосочетаний, содержащих слово w, выберем k наиболее частотных. Пусть Sum – сумма их частотностей. Тогда

$$SumK(w) = \frac{Sum}{K}$$

5.4.2. Признаки, полученные на основе выдачи глобальной поисковой машины

Для вычисления следующих двух признаков были использованы контексты употребления слов. В качестве таких контекстов мы используем сниппеты (аннотации документов в выдаче), получаемые от поисковой машины Яндекс через xml-интерфейс. Для вычисления признаков использовалось по 100 сниппетов из выдачи. Сниппеты, получаемые по одному запросу, соединяются в один документ и обрабатываются программой морфологического анализа. В результате для каждого набора сниппетов может быть определена совокупность лемм (слов в словарной форме) и их частотность встречаемости в данном наборе сниппетов. Для терминов существенным является принадлежность к предметной области. Простейший способ учесть фактор принадлежности к предметной области является задание списка маркеров предметной области, включающих некую совокупность (от нескольких единиц до нескольких десятков) наиболее характерных слов предметной области. Признак **Markers** учитывает количество таких слов, встретившихся в сниппетах, полученных для исходного слова. В данном случае в качестве маркеров мы использовали названия основных наук и образованных от них прилагательных: *математика, математический, физика, физический, химия, химический* и др.

Другим признаком, получаемым на основе сниппетов, является количество слов-определений в сниппете слова. Смысл признака **Neardefwords** (количество слов-определений в сниппетах) заключается в том, что если в сниппетах рядом с исходным словом встречаются слова,

характерные для определения в терминологических словарях (*это, тип, вид, класс* и др.), то, скорее всего, это термин, для которого вводится определение. Признак Neardefwords равен количеству таких слов, появившихся непосредственно рядом (слева или справа) с исходным словом в сниппетах, полученных по запросу, совпадающему с исходным словом.

5.4.3. Признак встречаемости слова в терминах тезауруса

Предположим, что разработка тезауруса предметной области уже начата, и в тезаурус внесена некоторая совокупность терминов. Тогда как дополнительный признак для определения терминологичности слова можно использовать признак количества терминологических словосочетаний, в которые входит данное слово – признак **FreqByThes.**

В текущем эксперименте мы использовали полную совокупность многословных терминов онтологии OEHT и, таким образом, пытались оценить, насколько можно предсказать терминологичность отдельного слова на этой основе.

5.4.4. Оценка качества извлечения терминологических слов

Оценка качества выделения однословных терминов проводилась для широкой области естественных наук. Все эксперименты проводились с выборкой величиной 5 тысяч слов, для которых были обсчитаны все вышеперечисленные признаки. В качестве эталонного множества терминов использовались однословные термины, включенные в состав Онтологии OEHT.

Табл. 5.1. Средняя точность для отдельных признаков слов

Признак	AvP
Частотность	46%
Tf.idf	51%

Признак	AvP
C-value	46%
Странность	52%
Наиболее частотное словосочетание Inside	51%
Sum3	52%
Sum10	54%
Sum50	54%
Близкие слова-определения NearDefWords	54%
Ключевые слова Markers	46%
Частотность по терминам FreqByThes	66%

Табл. 5.1 представляет характеристику средней точности AvP для отдельных характеристик слов. Отметим, что в качестве базового уровня, в котором не было сделано реально никакого разумного упорядочения, для эксперимента можно взять простое упорядочение по алфавиту, для которого величина средней точности оказалась равной 22%.

Как видно в табл. 5.1, такие признаки, как Tf.idf и Странность, которые учитывают контрастные коллекции, показали более хорошие значения средней точности по сравнению с простым признаком частотности, однако очевидно, что их использование не решает проблему определения терминологичности отдельных слов.

Самые высокие результаты по предсказанию однословных терминов показал признак частотности по многословным терминам FreqByThes. Предложенные нами признаки Sum10 и Sum50 показали самые высокие показатели средней точности среди признаков, полученных на коллекции документов.

Поскольку можно предположить, что вычисленные признаки отражают разные особенности однословных терминов, то является важным подобрать оптимальную комбинацию этих признаков. Для поиска наилучшей

комбинации были использованы алгоритмы машинного обучения. При этом выборка слов случайным образом разбивалась на две части (обучающая выборка и контрольная выборка) в соотношении 3 к 1.

Для подбора алгоритма комбинирования полученных признаков был использован программный пакет алгоритмов машинного обучения RapidMiner (www.rapidminer.com). Наилучшим методом по величине средней точности оказался метод логистической регрессии W-Logistic, на основе которого было достигнуто значение средней точности **AvP**=72%.

Таким образом, мы видим, что комбинация всех признаков дала результат по мере средней точности, почти на 40% превышающий наиболее известный способ упорядочения слов-кандидатов в термины tf.idf.

5.5. Алгоритм комбинирования признаков для извлечения двухсловных терминов

Для извлечения терминов-словосочетаний также был предложен набор различных факторов.

5.5.1. Признаки, полученные по коллекции документов предметной области

Частотность словосочетания. Признак частотности словосочетаний в коллекции **Freq** часто используется для извлечения терминологических словосочетаний, поскольку известно, что в число наиболее частотных словосочетаний коллекции предметной области входит достаточно высокая доля терминологических словосочетаний**.**

Взаимная информация слов MI словосочетания обычно вычисляется по следующей формуле:

$$MI(ab) = log\,(\frac{N \cdot freq\,(ab)}{freq\,(a) \cdot freq\,(b)})$$

где *freq* ()– частотность слов и словосочетаний в коллекции, *N* – число слов в коллекции. Признак показывает, насколько употребление слов в словосочетании отличается от их независимого употребления.

Кубическая взаимная информация слов MI_3 – модификация признака MI вычисляется по формуле следующего вида [40]:

$$MI_3(ab) = \log\left(\frac{N \cdot freq^3(ab)}{freq(a) \cdot freq(b)}\right)$$

Признак **усеченное словосочетание Inside** предназначен для выявления двухсловных словосочетаний, которые являются частью более длинного термина. Значение признака для словосочетания *ab* определяется следующим образом. Среди всех словосочетаний, извлеченных из коллекции документов, таких, что *ab* является частью этого (более длинного – три и более слов) словосочетания выбирается словосочетание {**ab**} с максимальной частотностью. Тогда

$$Inside(ab) = \frac{freq(*ab*)}{freq(ab)}$$

5.5.2. Признаки, полученные по сниппетам глобальной поисковой машины

Для получения сниппетов поисковой машине задавались запросы в виде самого словосочетания, а также запросы в виде его отдельных слов-компонентов. Например, при анализе словосочетания *инверсионная ось*, задаются поисковые запросы *инверсионная ось*, *инверсионная*, *ось*.

Признаки векторного сравнения сниппетов. Разработчики стандартов по созданию информационно-поисковых тезаурусов считают одним из важных факторов внесения в тезаурус таких словосочетаний предметной области, значения которых не следуют из значений их компонент. Мы предполагаем, что такая семантическая особенность

словосочетания может проявляться в контекстах употребления данного словосочетания.

Для сопоставления контекстов употребления словосочетания и составляющих его слов мы используем вектора S_{ab} лемм, полученных для словосочетания *ab*, и для его отдельных компонентов S_a, S_b. Сравнение векторов сниппетов производится с помощью вычисления скалярного произведения между векторами и фиксируется в признаках $Scalar_1$ и $Scalar_2$:

$$Scalar_1 = \frac{(S_{ab}, S_a)}{\|S_{ab}\| \cdot \|S_a\|}, \qquad Scalar_2 = \frac{(S_{ab}, S_b)}{\|S_{ab}\| \cdot \|S_b\|}$$

При замене частотностей лемм в векторах S_{ab} , S_a, S_b на булевские признаки {0, 1} в зависимости от присутствия или отсутствия леммы в сниппетах, получаются булевские вектора, и на их основе вычисляются соответствующие скалярные произведения – признаки $BinarScalar_1$ и $BinarScalar_2$. Признаки в виде бинарных скаляров показали высокую эффективность для извлечения устойчивых словосочетаний в [160].

Признаки максимально отличающегося контекста. Другим способом определения специфики употребления словосочетания является нахождение одного характерного слова, с которым чаще всего совместно встречается это словосочетание и относительно редко встречаются отдельные слова. Мы предполагаем, что если значение словосочетания не выводимо из значений его компонент, то это может проявиться в том, что это словосочетание употребляется в сниппетах рядом с такими словами, с которыми мало употребляются отдельные слова исходного словосочетания. При чем мы считаем, что словосочетание имеет тем большую семантическую особенность, чем больше максимальная разница между частотностью употребления некоторой леммы в сниппетах словосочетания по сравнению с употреблением этой же леммы в сниппетах отдельных слов.

Пусть лемма *L* встречается f_{ab} раз в сниппетах словосочетания S_{ab}, f_a – в сниппетах первого слова словосочетания S_a, f_b – в сниппетах второго слова

словосочетания S_b. Тогда коэффициент устойчивости SnipFreq_0 вычисляется по следующей формуле:

$$SnipFreq_0 = \max_L \cdot \left(f_{ab-a-b} \cdot \log\left(\frac{N - dlcol}{dlcol} \right) \right)$$

где $f_{ab\text{-}a\text{-}b} = max\ (f_{ab} - f_a - f_b,\ 0)$, $dlcol$ – частотность леммы в документах контрастной коллекции, N – количество документов в контрастной коллекции. Множитель $\log\left(\frac{N - dlcol}{dlcol} \right)$ idf-factor представляет собой известный в информационном поиске множитель *idf*, который помогает снизить влияние частотных общеупотребительных слов [130]. В качестве контрастной коллекции взята коллекция сайтов Белорусского Интернета, предоставленная компанией Яндекс в качестве экспериментальной коллекции врамках семинара по информационному поиску РОМИП.

В некоторых случаях двухсловное сочетание представляет собой несамостоятельный фрагмент более длинного словосочетания, и тогда такое двухсловное сочетание имеет наиболее высокую сочетаемость с остальными словами этого словосочетания. Для учета такой ситуации та же формула использовалась, чтобы посчитать наиболее характерное слово на расстоянии более, чем 1 и 2 слова от исходного словосочетания (признаки SnipFreq_1, SnipFreq_2).

Частотность упоминания словосочетания в собственных сниппетах. Признак частотности упоминания словосочетания в собственных сниппетах **FreqbySnip** может отражать различные особенности словосочетания. Если значение этого признака значительно меньше 100, то это означает, что поисковая машина не находит такое словосочетание в Интернет, и, таким образом, это словосочетание, возможно ошибочно извлечено при обработке коллекции (например, за счет неправильной лемматизации или неточной обработки таблиц в исходных документах). Если же значение этого признака значительно больше 100 (иногда этот признак

достигает величины 250-300 на 100 сниппетах), то это означает, что имеется множество контекстов, в которых это словосочетание подробно объясняется, является темой фрагмента, и, скорее всего, это словосочетание означает важное понятие или конкретную сущность.

Как и для случая выделения однословных терминов используются признаки «Количество слов-определений в сниппетах NearDefWords» и «Количество слов-маркеров предметной области Markers».

5.5.3. Признаки, полученные на основе лингвистической онтологии

На основе терминологического состава онтологии предметной области вычисляются признаки, которые должны помочь предсказать, относится ли к терминам данное словосочетание. Если словосочетание входит в состав терминов онтологии, то, естественно, это словосочетание исключается из множества терминов, являющихся базой для порождения признаков. Был протестирован ряд признаков, основанных на структуре и составе тезауруса, однако в настоящее время удалось эффективно использовать только три следующих признака.

Синоним к термину SynTerm. В текстах предметной области может встречаться много вариантов называния одного и того же термина, поэтому можно предположить, что если словосочетание похоже на словосочетание, которое уже считается термином (включено в тезаурус), то это словосочетание также является термином предметной области [17, 148].

Пусть слова *a* и *b* являются словами-компонентами словосочетания *ab*, по поводу которого нужно принять решение. Мы будем считать, что словосочетание *ab* является синонимом словосочетания *a′b′*, если *a* совпадает или является синонимом *a′*, а *b* совпадает или является синонимом *b′*. Синонимия отдельных слов задана в исходном тезаурусе посредством того, что слова указаны как текстовые входы одного и того же понятия тезауруса. Так, например, если слова *объект* и *предмет* указаны в тезаурусе как текстовые входы одного и того же понятия, то словосочетания *учебный*

объект и *учебный предмет* будут рассматриваться системой как потенциальные синонимы.

Синоним к нетермину SynNotTerm. Если на текущем этапе работы обнаружилось определенная в предыдущем разделе синонимичность данного словосочетания к словосочетанию, не включенному в начальный тезаурус, то эта информация фиксируется в специальном признаке.

Полнота описания Completeness. Для извлеченного словосочетания *ab* может оказаться, что его слова-компоненты *a* и/или *b* уже включены в начальный тезаурус в качестве текстовых входов понятий, и для соответствующих понятий уже описана некоторая совокупность отношений. Мы подсчитываем число всех отношений понятий, к которым возможно приписаны слова *a* и *b*, и фиксируем эту величину в виде характеристики Completeness. Предполагается, что, чем больше отношений у соответствующих понятий, тем более они важны для предметной области, и это может также поднять значимость словосочетания *ab*. Рассмотрим, например, словосочетание *собственный вектор*. Лемма *собственный* не соответствует ни одному понятию Онтологии ОЕНТ, а лемма *вектор* соответствует одному понятию ВЕКТОР, у которого 56 отношений. Таким образом, значение признака Completeness для этого словосочетания равно 56. Если у слов-компонентов словосочетания нет соответствующих понятий в тезаурусе, как, например, у словосочетания *последнее поступление*, то значение признака Completeness=0.

5.5.4. Оценка качества извлечения двухсловных терминов

Оценка качества извлечения двухсловных терминов проводилась в двух областях: области естественных наук и банковской области. В качестве эталонов использовались тезаурусы в соответствующих областях. Во экспериментых проводились на 5 тысячах самых частотных словосочетаний, извлеченных из соответствующих текстовых коллекций и имеющих

структуру типа прилагательное+существительное или существительное+ существительное в родительном падеже.

Для нахождения наилучшей комбинации признаков использовались методы машинного обучения, собранные в программном пакете RapidMiner (www.rapidminer.com). Для обучения наилучшей комбинации признаков исходная выборка делилась в соотношении 3 (обучающая выборка) к 1 (контрольная выборка). Качество извлечения терминов оценивалось с помощью меры средней точности AvP. В качестве базовых уровней, с которыми имеет смысл сравнивать упорядочение по совокупности признаков являюся упорядочение в алфавитном порядке (т.е. отсутствие всякого упорядочения по терминам) и упорядочение по частотности.

Было протестировано несколько методов машинного обучения из пакета RapidMiner, стабильно высокими были результаты метода логистической регрессии. Таблица 5.2 показывает значения AvP для отдельных признаков и их комбинации, полученных посредством метода логистической регрессии.

Два признака SynTerm и SynNotTerm являются бинарными, поэтому их некорректно оценивать с помощью оценки AvP. Признак SynTerm (Синоним к термину) очень хорошо "отделяет" (пользуясь терминологией логических алгоритмов классификации) термины, т. е. имеет место высоко информативная закономерность: если $SynTerm\ (x)=1$, то x – термин.

Из таблицы можно видеть, что использование одного и того же набора признаков и комбинирования методом машинного обучения позволяет получить значительно более высокие величины средней точности. Вместе с тем, видно, что в разных областях вклады отдельных признаков различаются. Так, например, в банковской области средняя точность для признака частотности Freq имеет максимальную величину, а этот же признак в ественно-научной области имеет относительно низкое значение.

Табл. 5.2 Значения средней точности для отдельных признаков и для комбинации признаков методом логистической регрессии.

Feature	AvP (Банковская сфера) %	AvP (Естественные науки)%
Alphabet	40%	57%
Frequency	57%	66%
MI	43%	64%
MI3	45%	67%
Inside	55%	75%
FreqBySnip	53%	69%
NearDefWords	49%	73%
$Scalar_1$	42%	61%
$Scalar_2$	45%	60%
$Boolean_1$	49%	64%
$Boolean_2$	48%	62%
$SnipFreq_0$	34%	66%
$SnipFreq_1$	38%	67%
$SnipFreq_2$	38%	67%
Markers	40%	65%
Completeness	52%	69%
SnipTitle	50%	-
Logistic Regression	**79% (+38.6% от Freq)**	**83% (+25.8% от Freq)**

Можно объяснить этот феномен относительной узостью банковской области. Документы, относящиеся к банковской деятельности, содержат множество терминов из соседних предметных областей таких, как экономика или политика. Такие термины, естественно, обладают всеми свойствами терминов за исключением собственно принадлежности к заданной

предметной области, и в итоге признак частотности становится одним из самых определяющих.

Также можно видеть серьезное различие во вкладах признаков $SnipFreq_i$ в научной области и в банковской области, где значения средней точности для этих признаков оказались очень низкими. На наш взгляд, это объясняется тем, что банковская сфера относится к области, регулируемой нормативными актами. Кроме того, имеется массовый повтор цитат из этих актов, что приводит к искажению величин $SnipFreq_i$.

Чтобы оценить значимость признаков была проведена процедура отбора признаков, которая позволяет отобрать минимальное количество признаков, позволяющее достичь практически того же качества выделения терминов. Для научной области такими признаками оказались $Boolean_1$, Completeness, FreqBySnip, Inside, MI, Neardefwords, SynTerm (AvP – 82%). Для банковской области отобранными признаками являются Completeness, FreqBySnip, MI, NearDefWords, $Scalar_1$, $SnipFreq_0$, SynTerm (AvP – 78%). Повторяющие признаки для обеих областей специально выделены. Можно видеть, что в обоих случаях среди выбранных признаков оказались признаки всех трех типов: полученные из специализированной текстовой коллекции, полученные на основе результатов поиска в Интернет, и полученные на основе соответствующего тезауруса.

Таким образом, предложенные признаки терминов и их комбинация позволяют значительно улучшить качество выделения терминов в двух различных предметных областях.

Заключение к главе 5

Эксперты при отборе терминов предметной области в терминологические словари, в информационно-поисковые тезаурусы руководствуются большим количеством разных правил, критериев. Таким образом, и автоматический отбор терминов должен базироваться на

многофакторных моделях. Кроме того, такие многофакторные модели должны быть переносимы с одной предметной области на другую.

В данной главе описан новый подход по выявлению большого количества признаков, относящихся к трем различным типам, для автоматического извлечения терминов из текстов и комбинирования этих признаков методами машинного обучения. Новизна предложенного подхода заключается и в том, что предложенный набор признаков нацелен на развитие существующей онтологии и позволяет настроиться на тип ресурса.

Полученная модель была протестирована на двух предметных областях, что позволило исследовать вопросы переносимости созданной модели. Модель может использоваться как на этапе начальной разработки терминологического ресурса, так и для его пополнения, поскольку важным учитываемым фактором являются признаки, основанные на уже отобранных терминах.

Заключение и основные результаты

Для того чтобы сделать обработку текстов для приложений информационного поиска более глубокой и более качественной необходимо создавать специализированные лингвистические ресурсы. На базе проведенного анализа существующих ресурсов и многочисленных экспериментов была разработана модель лингвистического ресурса для автоматической обработки текстов в приложениях информационного поиска.

Разработанная модель стала основой для разработки совокупности лингвистических ресурсов в ряде предметных областей, в том числе такие ресурсы, как Тезаурус русского языка РуТез и Онтология по естественным наукам и технологиям ОЕНТ.

Для применения разработанных лингвистических ресурсов в автоматической обработке текста был предложен и реализован ряд алгоритмов, которые были объединены в программно-лингвистический комплекс АЛОТ. Созданные лингвистические ресурсы и методы обработки текстов используются для обработки потоков документов в Университетской информационной системе РОССИЯ. Созданные технологии и ресурсы применяются в различных проектах с государственными и коммерческими организациями.

Таблица 6 представляет наиболее значимые проекты, в которых были использованы описанные алгоритмы и методы. В столбцах показан тип технологии, в том числе

- столбец «ЛО ОПТ» – означает поставку Общественно-политического тезауруса,
- столбец «Новые ЛО» означает создание новой лингвистической онтологии для предметной области заказчика,
- столбец «Извл. терминов» означает использование процедур извлечения терминов для терминологии предметной области заказчика,

- столбец «Поиск» означает создание информационно-поисковой системы с использованием концептуального индекса по ЛО,

- столбец «QA» означает создание вопросно-ответной системы,

- столбец «Рубрикация» – создание системы автоматической рубрикации,

- столбец «Аннотирование» – поставку системы автоматического аннотирование отдельных документов,

- столбец «Кластеризация» – использование концептуального индекса по ЛО как базы для кластеризации документов,

- столбец «Обзорное реферирование» – поставку системы автоматического аннотирования по тематически близким документам,

Табл. 6. Описанные алгоритмы и методы, использованные в наиболее значимых проектах.

Основные проекты	Годы	ЛО: ОПТ	Новые ЛО	Извл. терминов	Поиск	QA	Рубрикация	Аннотирование	Кластеризация	Обзорное реферирование	Аналитические отчеты
ГосДума ФС РФ	1999-н/в	✔			✔		✔	✔			
ЦБ РФ	2006-н/в	✔	✔	✔			✔	✔		✔	✔
ФСБ РФ	2000-н/в	✔	✔	✔	✔		✔	✔	✔	✔	✔
ГАС «Выборы» (ФКЗ «Право»)	1997-н/в	✔	✔	✔	✔	✔	✔				
НПП «Гарант-Сервис»	2002-н/в	✔				✔	✔	✔			
Рамблер. Новости	2008-н/в	✔					✔	✔	✔	✔	
Минюст РФ	2007	✔			✔						
Мин-во экологии МО	2007	✔	✔	✔	✔		✔	✔			✔
НИЦ «Квант»	2004	✔	✔	✔	✔		✔	✔			
Счетная палата РФ	2003			✔							
ИППИ РАН (Упр. спецпрограмм)	1996	✔			✔		✔	✔			

- столбец «Аналитические отчеты» - изготовление системы автоматической генерации аналитических отчетов по тематике заказчика; технология построения аналитических отчетов представляет собой сочетание технологий автоматической рубрикации и автоматического аннотирования.

Основными оригинальными результатами, полученными в диссертации, являются следующие:

1. Предложена новая формализованная модель базы знаний онтологического типа – лингвистической онтологии, предназначенной для использования в автоматической обработке текстов в широких предметных областях. Модель неоднократно использовалась для создания сверхбольших лингвистико-онтологических ресурсов в разных предметных областях.

2. Предложена новая модель представления тематической структуры текстов на основе согласованного учета свойств лексической и глобальной связности текста. Предложен и реализован алгоритм автоматического построения тематического представления содержания текстов, которое моделирует основное содержание текста посредством выделения тематических узлов – совокупностей близких по смыслу понятий текста.

3. Предложен и реализован метод концептуального индексирования документов для информационно-поисковой системы, базирующийся на понятиях лингвистической онтологии и тематическом представлении текста. В состав метода входит процедура автоматического разрешения лексической многозначности, основанная на информации о локальном и глобальном контексте употребления многозначного слова.

4. Предложен и реализован метод автоматического многошагового построения булевского выражения для длинного поискового запроса на естественном языке, включающий итерационное расширение запроса по отношениям лингвистической онтологии, подтвержденным поисковой выдачей.

5. Предложены и реализованы методы автоматической обработки текстов на основе концептуального индекса, включая:

- метод автоматической рубрикации документов, основанный на использовании тематического представления документов и описании рубрик в виде булевских выражений над понятиями лингвистической онтологии. На основе метода реализовано более 20 систем автоматической рубрикации;

- метод автоматического аннотирования отдельного документа и совокупности тематически близких документов на базе выделения из текстов наиболее содержательных предложений. В экспериментах показана высокая связность создаваемых аннотаций в сочетании с не уступающей другим методам полнотой представления информации.

Качество данных алгоритмов было экспериментально проверено в процессе независимой экспертизы в сравнении с другими методами на общественно доступных данных. Программы построения тематического представления текстов, порождения концептуального индекса, автоматической рубрикации и аннотирования объединены в единый программный комплекс тематического анализа текста.

6. Предложена многофакторная модель извлечения терминов предметной области. Реализованный в соответствии с предложенной моделью метод извлечения терминов основывается на вычислении для языковых выражений трех типов статистических характеристик и комбинировании их методами машинного обучения.

Список литературы

1. Advances in Automatic Text Summarization / Ed: I. Mani, Inderjeet, Maybury, Mark T. The MIT Press, 1999.

2. Ageev M., Dobrov B., Loukachevitch N. Text Categorization Tasks for Large Hierarchical Systems of Categories // In Proceedings of SIGIR-2002 Workshop on Operational Text Classification Systems / Eds. F.Sebastiani, S.Dumas, D.D.Lewis, T.Montgomery, I.Moulinier. Univ. of Tampere, 2002. P.49-52.

3. Ageev M., Dobrov B., Loukachevitch N. Sociopolitical Thesaurus in Concept-based Information Retrieval: Ad-hoc and Domain Specific Tasks // Cross-Language Evaluation Forum. Results of the CLEF 2005 Cross-Language System Evaluation Campaign / Eds.: C.Peters, V.Quochi. Springer Verlag, LNCS-4022, 2006. P. 141-150.

4. Agirre E., Rigau G. A Proposal for Word Sense Disambiguation using Conceptual Distance // In Proceedings of the First International Conference on Recent Advances in NLP. 1995.

5. Agirre E., Rigau G. Word Sense Disambiguation Using Conceptual Density // In Proceedings of COLING'96, Copenhagen, Danmark. 1996. P.16-22.

6. Agirre E., Magnini B., Lacalle O., Otegi A., Rigau G., Vossen P. SemEval–2007 Task 01: Evaluating WSD on Cross-Language Information Retrieval // In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), in conjunction with ACL. 2007.

7. AGROVOC Multilingual Agricultural Thesaurus. Fourth Edition. 1999.

8. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // In Proceedings of Eigth Text Retrieval Conference (Trec-8). 1999.

9. Aitchinson Y., Gilchrist A. Thesaurus construction: a practical manual. 2nd ed. L.:Aslib, 1987.

10. Ananiadou S. A methodology for automatic term recognition // In Proceedings of COLING-1994. 1994. P. 1034-1038.

11. Artale A., Franconi E., Guarino N., Pazzi L. Part-Whole Relations in Object-Centered Systems: An Overview // Data and Knowledge Engineering. 1996. V.20. P. 347-383.

12. Asmussen J., Pedersen B., Trap-Jensen L. DanNet: from Dictionary to WordNet // In Proceedings of GLDV-2007 Workshop: Lexical-Semantic and Ontological Resources. 2007. P. 1-10.

13. Atserias J., Climent S., Rigau G. Toward the Meaning Top Ontology: Sources of Ontological Meaning // In Proceedings of International conference Language Resources and Evaluation (LREC-2004). 2004. V.1. P. 11-14.

14. Barzilay R., Elhadad M. Using Lexical Chains for Text Summarization // ACL/EACL Workshop Intelligent Scalable Text Summarization. 1997.

15. Bentivogli L., Bocco A., Pianta E. ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge // In Proceedings of the Second Global WordNet Conference, Brno, Czech Republic. 2004. P. 39-46.

16. Blei D., Ng A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. N 3. P. 993-1022.

17. Bolshakova E. Recognition of Author's Scientific and Technical Terms // In: Computational Linguistics and Intelligent Text Processing / A. Gelbukh (Ed.). Lecture Notes in Computer Science, N 2004, Springer-Verlag. 2001.

18. Bouaud J., Bachimont B., Charlet J., Zweigenbaum P. Methodological principles for structuring an "ontology" // In Proceedings of IJCAI-95 Workshop "Basic Ontological Issues in Knowledge Sharing". 1995.

19. Brewster Ch., Iria J., Ciravegna F., Wilks Y. TheOntology: Chimaera or Pegasus // In Proceedings of Dagstuhl Seminar Machine Learning for the Semantc Web, 2005.

20. Brown G., Yule G. Discourse analysis. Cambridge University Press, 2001.

21. Brunn M., Chali Y., Pinchak C. Text Summarization Using Lexical Chains // In the Proceedings of the Document Understanding Conference (DUC-2001). 2001. P.135-140.

22. Budanitsky A. Lexical Semantic Relatedness and its application in Natural Language Processing. PhD Thesis. Technical Report CSRG-390, Computer Systems Research Group, University of Toronto. 1999.

23. Buenaga Rodriguez M., Gomez-Hidalgo J., Diaz-Agudo B. Using WordNet to complement training information in text categorization // In Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP 1997), Bulgaria. 1997. P. 150-157.

24. Buitellar P., Sacalenau B. Extending Synsets into Medical Terms. // In Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA. 2001.

25. Burgun A, Bodenreider O., Aubry M., Mosser J. Dependence relations in Gene Ontology: A preliminary study. // In Proceedings of Workshop on The Formal Architecture of the Gene Ontology, Leipzig, Germany. 2004.

26. Callan J.P., Croft W.B., Harding S.M. The INQUERY Retrieval System // A.M. Tjoa and I. Ramos (eds.), Database and Expert System Applications. Proceedings of {DEXA}-92, 3rd International Conference on Database and Expert Systems Applications. Springer Verlag, New York. 1992. P.78-93.

27. Cao G., Nie J., Bai J. Integrating Word Relationships into Language Models // In Proceedings of SIGIR-2005. 2005. P. 298-305.

28. Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // In Proceedings of the 21st Annual International ACM SIGIR Conference. 1998. P. 335-336.

29. Carlson L., Marcu D., Okurowski M. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory // Current directions in Discourse

and Dialog / Jan van Kuppevelt and Ronnie Smith, editors. Kluwer Academic Publishers, 2003. P. 85-112.

30. Carty J., Smeaton A. The Design of a Topic Tracking System // In the Proceedings of the 22nd Annual Colloquium on IR Research. 2000.

31. Chen H., Lynch K. J., Basu K., Ng T. D. Generating, integrating, and activating thesauri for concept-based document retrieval // IEEE Expert.1993. P. 25-34.

32. Gonzalo J., Chugur I., Verdejo F. Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index // In Proceedings of the SIGLEX Workshop on Word Senses and Multilinguality, in conjunction with ACL-2000, Hong Kong, China. 2000.

33. Chugur I., Gonzalo J., Verdejo F. Polysemy and sense proximity in the Senseval-2 Test Suite // In Proceedings of the ACL-2002 Workshop on "Word sense Disambigation:recent successes and future directions". 2002.

34. Clark P., Fellbaum Ch., Hobbs J. Using and Extending WordNet to Support Question-Answering // In Proceedings of the Fourth Global WordNet Conference (GWC'08), Hungary: University of Szeged, 2008. P. 111-119.

35. Climent S., Rodriguez H., Gonzalo J., Definitions of the links and subsets for nouns of the EuroWordNet project. Deliverable D005, WP3.1, EuroWordNet, LE2-4003. 1996.

36. Corcho O., Gomez-Perez A. A Roadmap to Ontology Specification Languages // Knowledge Engineering and Knowledge Management. Methods, Models and Tools. / Rose Dieng and Oliver Corby (eds). Springer, 2000. P. 80-96.

37. Cristea D., Ide N., Romary L. Veins Theory: A Model of Global Discourse Cohesion and Coherence // In Proceedings of Seventeenth Conference of Computational Lingustics (COLING 1998). 1998. P. 281-285.

38. Cruse D. Lexical Semantics. Cambridge, University Press. 1986.

39. Cyc Ontology Guide: Introduction. *(http://www.cyc.com/cyc-2-1/intro-public.html).*

40. Daille B., Gaussier E., Lang J.M. An evaluation of statistics scores for word association // In: Tbilisi Symposium on Logic, Language and Computation. CSLI Publications. 1998. P. 177-188.

41. Dang H.T., Overview of DUC 2006. National Institute of Standards and Technology (NIST). 2006. http://www-nlpir.nist.gov/projects/duc/pubs/2006papers/duc2006.pdf

42. Dijk van T. Semantic discourse analysis // Handbook of Discourse Analysis / Teun A. van Dijk, (Ed.), vol. 2. London: Academic Press. 1985. P. 103-136.

43. Dobrov B., Loukachevitch N., Nevzorova O., Fedunov B. Methods of automated design of application ontology // Journal of Computer and systems sciences international. 2004. V. 43. I. 2. P. 213-222.

44. Dobrov B., Loukachevitch N. Multiple Evidence for Term Extraction in Broad Domains // In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP-2011). 2011. P. 710-715.

45. Doran W., Stokes N., Carty J., Dunnion J. Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization //In Proceedings of CICLING-2004. 2004. P. 627-635.

46. Dumais S., Platt J, Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization // In Proceedings of International Conference on Information and Knowledge Management. 1998. P. 148-155.

47. Dumais S., Lewis D., Sebastiani F. Report on the Workshop on Operational Text Classification Systems (OTC-02) // In Proceedings of SIGIR-2002, Tampere, Finland. 2002.

48. Felber H. Terminology Manual. Unesco, Infoterm, 1984. 426 p.

49. Fine K. Ontological Dependence // In Proceedings of the Aristotelian Society 95. 1995. P. 269-290.

50. Foo Merkel. Using machine learning to perform automatic term recognition // In Proceedings of. LREC2010 Aquisition Workshop, Valetta, Malta. 2010.

51. Fox M.S., Gruninger M. On Ontologies and Enterprise modeling // In Proceedings of International Conference “Enterprise Integration Modeling Technology. 1997.

52. Galley M., McKeown K. Improving word sense disambiguation in lexical chaining // In Proceedings of IJCAI-2003. 2003.

53. Gangemi A., Guarino N., Oltramari A. Conceptual analysis of lexical taxonomies: the case of wordnet top-level // In Proceedings of the international conference on Formal Ontology in Information Systems. ACM Press, 2001.

54. Gangemi A., Navigli R., Velardi P. The OntoWordNet project: extension and axiomatisation of conceptual relations in Wordnet // In Proceedings of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE), Catania (Italy). 2003.

55. Gene Ontology. An Introduction to Gene Ontology. Код доступа: http://www.geneontology.org/GO.doc.shtml.

56. Gerstl P., Pribennow S. A conceptual theory of part-whole relations and its applications // Data and Knowledge Engineering. 1996. V.20. P. 305-322.

57. Grenon P. Spatio-temporality in Basic Formal Ontology: SNAP and SPAN, Upper-Level Ontology, and Framework for Formalization: PART I // IFOMIS Report 05/2003, Institute for Formal Ontology and Medical Information Science (IFOMIS), University of Leipzig, Leipzig, Germany. 2003.

58. Griffiths T., Steyvers M. A probalistic approach to semantic representation // In Proceedings of the 24th Annual Conference on the Cognitive Science Society. 2002.

59. Griffiths T., Steyvers M. Finding Scientific Topics // In Proceedings of the National Academy of Science. 2004. V.101. P. 5228-5235.

60. Gomez-Perez A., Fernandez-Lopez M., Corcho O. OntoWeb. Technical Roadmap. D.1.1.2. - IST project IST-2000-29243, 2001. (http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf)

61. Gomez-Perez A., Benjamins V.R. Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods // In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden. 1999.

62. Gomez-Perez A., Corcho O, Fernandez-Lopez M. Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing). Springer-Verlag, 2004.

63. Gonzalo J., Verdejo F., Chugur I., Cigarrán J. Indexing with WordNet synsets can improve text retrieval // In Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP. 1998.

64. Gruber T.R. A translation approach to portable ontologies // Knowledge Acquisition. 1993. V. 5(2). P. 199-220.

65. Guarino N. Concepts, attributes and arbitrary relations // Data Knowledge Engineering. 1992. V.8. P. 249-261.

66. Guarino N., Giaretta P. Ontologies and Knowledge Bases: Towards a Terminological Clarification // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing / N. Mars (ed.). Amsterdam: IOS Press, 1995. P. 25-32.

67. Guarino N. Formal Ontology and Information Systems // In Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98 /N. Guarino, editor, Trento, Italy, IOS Press. 1998. P. 3-15.

68. Guarino N. Some Ontological Principles for Designing Upper Level Lexical Resources // In Proceedings of First International Conference on Language Resources and Evaluation, Granada, Spain. 1998.

69. Guarino N., Welty C. Evaluating ontological decisions with ONTOCLEAN // Communications of the ACM. 2002. V. 45(2). P. 61-65.

69a. Guizzardi G. Ontological foundations for structural conceptual models. CTIT-PhD-thesis Series No 05-74, 2005.

70. Halliday M., Hasan R. Cohesion in English. London: Longman, 1976.

71. Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V., Morarescu P. FALCON: Boosting Knowledge for Answer Engines // In Proceedings of TREC-9. 2000.

72. Harnly A., Nenkova A., Passonneau R. Rambow O. Automation of summary evaluation by the pyramid method // In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'2005), Borovets, Bulgaria, 2005.

73. Hasan R. Coherence and Cohesive harmony // Understanding reading comprehension / J. Flood, editor.Newark, DE: IRA, 1984. P. 181-219.

74. Hayes Ph. Intelligent High-Volume Processing Using Shallow, Domain-Specific Techniques // Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval. New Jersey, 1992. P. 227-242.

75. Hepp M. Possible Ontologies: How Reality Constrains the Development of Relevant Ontologies // IEEE Internet Computing. 2007.Vol. 11, No. 1. P. 90-96.

76. Hirst G., St-Onge D. Lexical Chains as representation of context for the detection and correction malapropisms // WordNet: An electronic lexical database and some of its applications /C. Fellbaum, editor. Cambrige, MA: The MIT Press, 1998.

77. Hirst G. Ontology and the Lexicon // Handbook on Ontologies in Information Systems. Berlin: Springer, 2003.

78. Hirst G., Morris J. The subjectivity of Lexical Cohesion in Text // Computing attitude and affect in text / In James C. Chanahan, Yan Qu, and Janyce Wiebe, editors. 2005. P. 41-48.

79. Hlava M., Hainebach R. Multilingual Machine Indexing // In Proceedings of The Ninth International Conference on New Information Technology. 1996. P. 105-120.

80. Hollingsworth W., Teufel. S. Human Annotation of Lexical Chains // In Workshop proceedings ``ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications'', SIGIR-2005, Salvador, Brazil. 2005.

81. Hotho A., Bloehdorn S. Boosting for Text Classification with semantic features // In Proceedings of the Workshop on Mining for and from the Semantic Web at the 10th International Conference on Knowledge Discovery and Data Mining (KDD-2004). 2004. P.70-87.

82. Hovy E., Hermjakob U., Lin C.-Y. The use of external knowledge in factoid QA // In Proceeding 10th Text Retrieval Conference (TREC 2001). 2001.

83. ISO 2788-1986. Guidelines for the establishment and development of monolingual thesauri. 1986.

84. ISO 5964-1985. Guidelines for the establishment and development of multilingual thesauri. 1985.

85. Jensen L., Martinez T. Improving text classification by using coceptual and contextual features // In Proceedings of the Workshop on Text Mining at the 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining (KDD 00). 2000. P. 101-102.

86. Jeon J., Croft B., Lee J.H. Finding Similar Questions in Large Question and Answer Archives // In Proceedings CIKM-2005. 2005. P. 84-90.

87. Jiang J., Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy // In Proceedings of COLING-1997. 1997.

88. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features // In Proceedings of ECML-98, 10th European Conference on Machine Learning. 1998.

89. Kehagias A., Petridis V., Kaburlasos V., Fragkou P. A comparison of word- and sense-based text classification using several classification algorithms // Journal of Intelligent Information Systems. 2003. V. 21(3). P. 227-247.

90. Kilgarriff A., Rosenzweig J. Framework and Results for English Senseval // Computers and the Humanities. 2000. V34. P. 15-48.

91. Kluck M. GIRT Data in the Evaluation of CLIR Systems – from 1997 until 2003 // In Proceedings of Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003 /C. Peters (ed). LNCS-3237. Springer, 2003.

92. Kumar A., Smith B. The ontology of blood pressure: a case study in creating ontological partitions in biomedicine. 2004.

93. Kupiec J. MURAX: a Robust Linguistic Approach for Question Answering Using Om-line Encyclopedia // In Proceedings of SIGIR-1993. 1993. P. 181-190.

94. Landes S., Leacock C., Tengi, R.I. Building semantic concordances // WordNet: An Electronic Lexical Database / Fellbaum, C. (ed.). Cambridge (Mass.): The MIT Press, 1998.

95. Landauer T., Dumais S. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knwoledge // Psychological Review. 1997. V. 104. P. 211-240.

96. Leacock C., Chodorow M. Combining local context and WordNet similarity for word sense identification // WordNet: An electronic lexical database / Fellbaum, C. (ed.). Cambridge (Mass.): The MIT Press, 1998.

97. Lenat D., Miller G., Yokoi T. CYC, WordNet, and EDR: critiques and responses // Communications of the ACM. 1995. V. 38 , N 11. 1995. P. 45-48.

98. Lesk M. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone // In Proceedings of the 5th annual international conference on Systems documentation SIGDOC. 1986. P. 24-26.

99. Lewis D. Applying Support Vector Machines to the TREC-2001 Batch Filtering and Routing Tasks // In Proceedings of TREC-2001. 2001.

100. Li J., Sun L., Kit C., Webster J. A Query-Focused Multi-Document Summarizer Based on Lexical Chains // In Proceedings of the Document Understanding Conference DUC-2007. 2007.

101. Li S., Ouyang Y., Sun B., Guo Z. IBM “Peking University at DUC 2006” // In Proceedings of DUC-2006. 2006.

102. Liang A., Lauser B., Sini M., Keizer J., Katz S. From AGROVOC to the agricultural ontology service/concept server: An OWL model for managing ontologies in the agricultural domain // In Proceedings of OWL: Experiences and Directions Workshop. 2006.

103. Liddy E.D., Diekema A.R., Yilmazel O., Chen J., Harwell S., He, L. Finding Answers to Complex Questions // New Directions in Question Answering / Maybury, M. (Ed.). 2004. P. 141-152.

104. Lin Chin-Yew. ROUGE: a Package for Automatic Evaluation of Summaries // In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain. 2004.

105. Liu Sh., Liu F., Yu C., Meng W. An effective approach to document retrieval via utilizing WordNet and recognizing phrases // In Proceedings of SIGIR-2004. 2004. P. 266-272.

106. LIV (Legislative Indexing Vocabulary*)*. Congressional Research Service. The Library of Congress. Twenty-first Edition. 1994.

107. Loebe F. Abstract vs. Social Roles: A Refined Top-level Ontological Analysis // In Proceedings of the 2005 AAAI Fall Symposium 'Roles, an Interdisciplinary Perspective: Ontologies, Languages, and Multiagent Systems / Guido Boella, James Odell, Leendert van der Torre and Harko Verhagen (ed.). AAAI Press, 2005. P.93-100.

108. Loukachevitch N., Dobrov B. Thesaurus-Based Structural Thematic Summary in Multilingual Information Systems // Machine Translation Review. 2000. N 11. P. 10-20.

109. Loukachevitch N., Dobrov B. Thesaurus as a Tool for Automatic Detection of Lexical Cohesion in Texts // In Proceedings of 5th JADT. 2000. P. 155-162.

110. Loukachevitch N.V., Dobrov B.V. Development and Use of Thesaurus of Russian Language RuThes // In Proceedings of workshop on WordNet

Structures and Standartisation, and How These Affect WordNet Applications and Evaluation. (LREC2002) / Dimitris N. Christodoulakis. 2002. P. 65-70.

111. Loukachevitch N., Dobrov B., Ageev M. Text Categorization Tasks for Large Hierarchical Systems of Categories // In SIGIR 2002 Workshop on Operational Text Classification Systems / Eds. F.Sebastiani, S.Dumas, D.D.Lewis, T.Montgomery, I.Moulinier. Univ. of Tampere, 2002. P. 49-52.

112. Loukachevitch N., Dobrov B. Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains // In Proceedings of Second International WordNet Conference GWC-2004. 2004. P.163-168.

113. Loukachevitch N., Dobrov B. Development of Ontologies with Minimal Set of Conceptual Relations // In Proceedings of Fourth International Conference on Language Resources and Evaluation / Eds: M.T.Lino et al., vol. VI. 2004. P.1889-1892.

114. Loukachevitch N., Dobrov B. Development of Bilingual Domain-Specific Ontology for Automatic Conceptual Indexing // In Proceedings of Fourth International Conference on Language Resources and Evaluation / Eds: M.T.Lino et al., vol. VI. 2004. P. 1993-1996.

115. Loukachevitch N., Dobrov B. Ontological Types of Associative Relations in Information Retrieval Thesauri and Automatic Query Expansion // In Proceedings of OntoLex 2004: Ontologies and Lexical Resources in Distributed Environments / Eds: A.Oltramari et al. 2004. P. 24-29.

116. Loukachevitch N., Dobrov B. Combining Evidence for Automatic Extraction of Terms // In Procedings of 4th International conference on Pattern Recognition and Machine Intelligence. Lecture Notes in Computer Science, V. 6744. Springer Verlag, 2011. P. 235-241.

117. Loukachevitch N. Concept Formation in Linguistic Ontologies. Conceptual Structures: Leveraging Semantic Technologies // In Proceedings of ICCS-2009 / Eds Sebasian Rudolph, Frithjof Dau, Sergei O. Kuznetsov. Springer Verlag, 2009. LNAI-5662. P. 2-22.

118. Loukachevitch N. Establishment of taxonomic relationships in linguistic ontologies // Knowledge processing and data analysis. Springer Verlag, 2011. LNCS-6581. P. 232-242.

119. Lowe E.J. Ontological dependence // Stanford encyclopedia of Philosophy. (http://plato.stanford.edu/entries/dependence-ontological/)

120. Lukashevich N.V. Concepts in formal and linguistic ontologies // Automatic Documentation and Mathematical Linguistics. 2011. V45. N 4. P. 155-162.

121. Maedche A., Staab S. Learning Ontologies for the Semantic Web // In Proceedings of Semantic Web Workshop, Hongkong. 2001.

122. Maedche A., Zacharias V. Clustering Ontology-based Metadata in the Semantic Web // In Proceedings PKKD-2002. Helsinky, 2002. P. 342-360.

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

124. Magnini B., Speranza M. Merging Global and Specialized Linguistic Ontologies // In Proceedings of OntoLex-2002. 2002.

125. Mahesh K., Nirenburg S. A Situated Ontology for Practical NLP // In Procedings of Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada. 1995.

126. Mani I., House D., Klein G., Hirshman L., Firmin Th., Sundheim B. The TIPSTER SUMMAC Text Summarization Evaluation // In Proceeding of EACL-99. 1999. P. 77-85.

127. Mani I., House D., Klein G., Hirshman L., Firmin Th., Sundheim B. SUMMAC: a text summarization evaluation // Natural Language Engineering. 2002. V.8, N 01. P. 43-68.

128. Mann W.C., Thompson S.A. Rhetorical Structure Theory: Description and Construction of Text Structures // Natural Language Generation. 1987.

129. Manning Ch., Raghavan P., Shutze H. Introduction to Information Retrieval. Cambridge University Press, 2008.

130. Mansuy T., Hilderman R. A characterization of WordNet Features in Boolean Models for Text Categorization // In Proceedings of Australasian Data Mining Conference (AusDM-2006). 2006. V. 61. P. 103-109.

131. Marinelli R., Tiberi M., Bindi R. Encoding Terms from a Scientific Domain in a Terminological Database: Methodology and Criteria // In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 2008.

132. Margolis E., Laurence S. Concepts. Stanford Encyclopedia of Philosophy. – 2006. Код доступа http://plato.stanford.edu/entries/concepts/#ClaThe .

133. Masolo C., Vieu L., Bottazzi E. Catenacci C., Ferrario R., Gangemi A., Guarino N. Social roles and their descriptions // In Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning. AAAI Press. 2004.

134. Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., Shneider L. WonderWeb. Final Report. Deliverable D18. 2003.

135. Mauldin M. Retrieval performance in Ferret a conceptual information retrieval system // In Proceedings of 14th SIGIR Conference. 1991. P. 347-355.

136. Medelyan O. Computing Lexical Chains with Graph Clustering // In Proceedings of the ACL 2007 Student Research Workshop. 2007. P. 85-90.

137. Mihalcea R. Co-training and Self-training for Word Sense Disambiguation // In Proceedings of CoNLL-2004. 2004.

138. Miller G. Nouns in WordNet // WordNet – An Electronic Lexical Database / Fellbaum, C (ed). The MIT Press, 1998. P. 23-47.

139. Miller G., Fellbaum C. Morphosemantic links in WordNet // Traitement automatique de langue, 44.2. 2003. P. 69-80.

140. Min-Yen Kan, Klavans J., McKeown K. Linear Segmentation and Segment Relevance // In Proceedings of 6th International Workshop of Very Large Corpora (WVLC- 6). 1998. P. 197-205.

141. Mochizuki H., Iwayama M., Okumura M. Passage Level Document Retrieval Using Lexical Chains // RIAO 2000, Content Based MultiMedia Information Access. 2000. P. 491-506.

142. Moldovan D., Novischi A. Lexical Chains for Question Answering // In Proceedings of International Conference on Computational Linguistics (COLING-2002). 2002. P. 674-680.

143. Molla D., Vicedo J. Question Answering in Restricted Domains: An Overview // Journal of Computational linguistics. 2007. V. 33, N 1. 2007. P. 41-61.

144. Montejo-Ráez A., Steinberger R., López A. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections // Advances in Natural Language Processing: 4th International Conference, EsTAL-2004. Springer, 2004. P. 1-12.

145. Morris J., Hirst G. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text // Computational Linguistics, 17(1). 1991. P. 21-45.

146. Morris J., Beghtol, C., Hirst G. Term relationships and their contribution to text semantics and information literacy through lexical cohesion // In Proceedings 31st Annual Conference of the Canadian Association for Information Science, Halifax, Canada. 2003

147. Motschnig-Pitrik R., Kaasboll J. Part-Whole Relationship Categories and their Application in Object-Oriented Analysis // IEEE TSE. 1999. V. 11(5). P.779-797.

148. Nenadic G., Ananiadou S., McNaught J. Enhancing automatic term recognition through recognition of variation // In Proceedings of 20th International Conference on Computational Linguistics (COLING-2004). 2004. P. 604-610.

149. Nenkova A., Louis A. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization // In Proceedings of ACL-08. 2008. P. 825-833.

150. Niles I., Pease A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology // In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. 2003. P.412-416.

151. Nirenburg S., Wilks Y., What's in a symbol: Ontology, representation, and language // Journal of Experimental and Theoretical Artificial Intelligence. 2001. V. 13(1). P. 9-23.

152. Nirenburg S., Raskin V. Ontological Semantics. MIT Press, 2004.

153. Nolt J. Free logic // Stanford Encyclopedia of Philosophy, 2010. http://plato.stanford.edu/entries/logic-free/#1.2

154. Noy N.F., McGuinness D. Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. 2001. Рус. Перевод: *Разработка онтологий 101: руководство по созданию Вашей первой онтологии (http://ifets.ieee.org/russian/depository/ontology101_rus.doc).*

155. Noy N., Wallace E. Simple part-whole relations in OWL Ontologies. W3C Technical report. 2005. (http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/index.htm)

156. Obrst L. Ontologies for Semantically Interoperable Systems // In Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM-2003). 2003. P. 366-369.

157. Ogilvie P., Callan J. Experiments using the Lemur Toolkit // In Proceedings of the 2001 Text REtrieval Conference. 2002.

158. Page L., Brin S., Motwani R., Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, 1999.

159. Pazienza M., Stellato A. Linguistic Enrichment of Ontologies: a Methodological Framework // In Proceedings Ontolex-2006 Workshop. 2006.

160. Pecina P., Schlesinger P. Combining association measures for collocation extraction // In Proceedings of Annual Meeting of the Association for Computational Linguistics, ACL-2006. 2006.

161. Pedersen T., Patwardhan S., Michelizzi J. WordNet: Similarity-measuring the relatedness of concepts // In Proceedings of the 19th National Conference on Artificial Intelligence, AAAI-2004. 2004. P. 144-152.

162. Peter H., Sach H., Bechstein C. Smartindexer – Amalagamating Ontologies and Lexical Resources for document indexing // In Proceedings of OntoLex-2006. 2006.

163. Peters W., Peters I., Vossen P. Automatic sense clustering in EuroWordNet // In Proceedings of the 1st. International conference on Language Resources and evaluations. 2000.

164. Petras V. GIRT and the Use of Subject Metadata for Retrieval // In Multilingual Information Access for Text, Speech and Images. 5th workshop of the Cross-language Evaluation Forum, CLEF-2004. LNCS-3491. Springer-Verlag, 2004. P. 298-309.

165. Petras V. How One Word Can Make all the Difference – Using Subject Metadata for Automatic Query Expansion and Reformulation // In Multilingual Information Access for Text, Speech and Images. 6th workshop of the Cross-language Evaluation Forum, CLEF-2005. Springer-Verlag, 2005.

166. Ponte J., Croft B. A Language Modeling Approach to Information Retrieval // In Proceedings of SIGIR-1998. 1998. P. 275-281.

167. Pouliquen B., Steinberger R., Ignat C. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus // In Proceedings of the International Conference *Recent Advances in Natural Language Processing,* Borovets, Bulgaria. 2003. P. 401-408

168. Radev D., McKeown K., Hovy E. Introduction to the Special Issue on Summarization // Computational linguistics. 2002. P. 399-408.

169. Reed S., Lenat D. Mapping ontologies into Cyc // In Proceedings of AAAI 2002 Conference Workshop on Ontologies for the Semantic Web, Edmonton, Canada. 2002.

170. Reeve L., Han H., Brooks A. BioChain: Using Lexical Chaining for Biomedical Text Summarization // In Proceedings of the ACM Symposium on Applied Computing. 2006. P. 180-184.

171. Resnik P. Using information content to evaluate semantic similarity // In Proceedings of IJCAI-1995. 1995.

172. Riloff E., Lehnert W. Information extraction as a basis for high-precision text classification // ACM Transactions on Information Systems. 1994. V. 12, N 3. P. 296-333.

173. Robertson S., Walker S., Hancock-Beaulieu M., Gatford M. Okapi in Trec-3 // In Proceedings of Text Retrieval Conference TREC-3. NIST Special publication 500-225. 1994. P. 109-126.

174. Rose T, Stevenson M., Whitehead M. The Reuters Corpus Volume 1 – from Yesterday News to tomorrow's Language // In Proceedings of the Third International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria. 2002.

175. Roventini A., Marinelli R. Extending the Italian WordNet with the Specialized Language of the Maritime Domain // In Proceedings of Second International WordNet Conference GWC-2004. 2004. P. 193-198.

176. Sagri M., Tiscornia D., Bertagna F. Jur-WordNet // In Proceedings of Second International WordNet Conference GWC-2004. 2004. P. 305-310.

177. Salton G. Automatic Text Processing - The Analysis, Transformation and Retrieval of Information by Computer. Addison-Wesley, Reading, MA, 1989.

178. Sanderson M. Word Sense Disambiguation and information retrieval // In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994.

179. Scott S., Matwin S. Text classification using WordNet hypernyms // In Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems (Coling-ACL 1998), Montreal, Canada. 1998. P.45-52.

180. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Reviews. 2002. V.34, N 1.

181. Shah Ch., Croft B. Evaluating High Accurace Retrieval Techniques // In Proceedings of SIGIR-2004. 2004. P. 2-9.

182. Schott H. Thesaurus for Social Sciences. 2 vols. Vol.1. German – English. 2. English – German. Bonn: Informations-Zentrum Sozialwissenshaften. 2000.

183. Silber G., McCoy K. Efficiently computed lexical chains as an intermediate representation for automatic text summarization // Computational Linguistics. 2003. V.29, N 1.

184. Simons P. Parts. A study in Ontology. Oxford University Press, 1987.

185. Smith B. Basic tools of formal ontology // Formal Ontology in Information Systems / N. Guarino (ed.). 1998.

186. Smith B. Beyond Concepts: Ontology as Reality Representation // In Proceedings of International Conference on Formal Ontology and Information Systems FOIS-2004. 2004.

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

188. Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. Reengineering Thesauri for New Applications: the AGROVOC Example. Article No. 257, 2004-03-17. 2004.

189. Song F., Croft B. A General Language Models for Information Retrieval // Research and Development in Information Retrieval. 1999. P. 279-280.

190. Sowa J. Using a Lexicon of Canonical Graphs in a semantic interpreter // Relational models of lexicon / M.Evens. Cambridge University press. 1988. P.113-137.

191. Sowa J. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA. 2000.

192. Sowa J. Building, Sharing and Merging Ontologies. Режим доступа: http://www.jfsowa.com/ontology/ontoshar.htm

193. Srikanh M., Srikanh R. Biterm language models for document retrieval // In Proceedings SIGIR-2002. 2002. P. 425-426.

194. Stairmand M. Textual content analysis for information retrieval // In Proceedings of 20th Annual ACM SIGIR Conference (SIGIR-97). 1997. P. 140-147.

195. Steinberger R., Hagman J. Scheer St. Using Thesauri for Automatic Indexing and Visualisation // In Proceedings of OntoLex-2000. 2000. P. 130-141.

196. Steinmann F. The representation of roles in object-oriented and conceptual modelling // Data and Knowledge engineering. 2000. V. 35, N 1. P. 83-106.

197. Stokes N., Hatch P., Carthy J. Lexical semantic relatedness and online news event detection // In Proceedings of the Annual 23rd ACM SIGIR Conference on Research and Development (SIGIR-00). 2000. P. 324-325.

198. Stokes N., Carthy J, Smeaton A.F. SeLeCT: A lexical Cohesion based News Story Segmentation System // Journal of AI Ccommunications. 2004. V. 17, N 1. P. 3-12.

199. Tomlin R. S., Forrest L., Pu M. M. Discourse semantics // Discourse as structure and process / T. van Dijk (Ed.). London: Sage, 1997. P. 63-111.

200. Tsujii J., Ananiadou S. Thesaurus or logical ontology, which one do we need for text mining? // Language Resources and Evaluation. 2005. V. 39, N 1. P. 77-90.

201. Tudhope D., Alani H., Jones Cr. Augmenting Thesaurus Relationships: Possibilities for Retrieval // Journal of Digital Libraries. 2001. V.1, N 8.

202. Tudhope D., Taylor C. Navigation via Similarity: automatic linking based on semantic closeness // Information Processing and Management. 1997. V. 33, N 2. P. 233-242.

203. Turdakov D., Lizorkin D. HMM Expanded to Multiple Interleaved Chains as a Model for Word Sense Disambiguation // In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computations. 2009. P. 549-559.

204. Varzi A. A Note on Transitivity of Parthood // Applied Ontology. 2006. V. 1, N 2. P. 141-146.

205. Vechtomova O., Jones R., Dias G. Report on the ACM International Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-World Applications // Sigir Forum. 2005. V 39, N2. P. 42-45.

206. Vivaldi J., Marquez L., Rodriguez H. Improving Term Extraction by System Combination Using Boosting // In Proceedings of ICML-2001. 2001. LNCS-2167. P. 515-526.

207. Voorhees E. Query expansion using lexical-semantic relations // In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994. P. 61-69.

208. Voorhees E. Using WordNet for Text Retrieval // WordNet – an Electronic Lexical Database. MIT Press, 1998. P. 285-304.

209. Voorhees E. Natural Language Processing and Information Retrieval // Information Extraction: Towards Scalable, Adaptable Systems / M.T. Pazienza (ed.). New York: Springer, 1999. P. 32-48.

210. Voorhees E. Overview of the TREC 2004 Question Answering Track. NIST Special Publication 500-261. 2004.

211. Vossen P. (ed.). EuroWordNet: A multilingual Database with Lexical Semantic Network. Dodrecht, 1998.

212. Vossen P. Wordnet, EuroWordNet and Global WordNet // International Conference RANLP-2003, Borovets, Bulgaria. 2003.

213. Vossen P., Glaser E., Gradinaru M., van Steenwijk R., van Zutphen H. Meaning-full Effects on Information Retrieval. IST-2001-34460, Deliverable 8.3. 2005.

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

215. Wasson M. Classification Technology at LexisNexis // In Proceedings of SIGIR 2001 Workshop on Operational Text Classification. 2001.

216. Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F. Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. 1999.

217. Wilks Y. Ontotherapy: or how to stop worrying about what there is // Ontolex 2002, Workshop on Ontologies and Lexical Knowledge Bases, LREC-2002. 2002.

218. Wilks Yorick. The Semantic Web as the apotheosis of annotation, but what are its semantics? // IEEE Intelligent Systems. 2008.

219. Will L. Thesaurus consultancy // The thesaurus: review, renaissance and revision / Sandra K. Roe and Alan R. Thomas, editors. New York, London : Haworth, 2004. 209p.

220. Winston M., Chaffin R., Herrmann D. A Taxonomy of Part-Whole Relations // Cognitive Science, 11. 1987. P. 417-444.

221. Woods W. Conceptual Indexing: A Better Way to Organize Knowledge. Sun Microsystems, Inc., Technical Report: TR-97-61. 1997.

222. Yang Y., Liu X. A re-examination of text categorization methods // In Proceedings of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99). 1999. P. 42-49.

223. Z39.19 – Guidelines for the Construction, Format and Management of Monolingual Thesauri. NISO, 1993.

224. Zhai C., Lafferty J. A study of smoothing methods for language models applied to ad hoc information retrieval // In Proceedings of SIGIR-2001. 2001. P. 334-342.

225. Zhang Z., Iria J., Brewster Ch., Ciravegna F. A Comparative Evaluation of Term Recognition Algorithms // In Proceedings of Sixth International Language Resources and Evaluation (LREC-08). 2008.

226. Агеев М.С., Добров Б.В., Журавлев С.В., Лукашевич Н.В., Сидоров А.В., Юдина Т.Н. Технологические аспекты организации доступа к разнородным информационным ресурсам в университетской информационной системе Россия // «Электронные библиотеки». 2002. т.5, Вып.2.

227. Агеев М.С., Кураленок И.Е. Официальные метрики РОМИП'2004. // Российский семинар по Оценке Методов Информационного Поиска. Пущино, 2004.

228. Агеев М., Добров Б., Красильников П., Лукашевич Н., Павлов А., Сидоров А., Штернов С. УИС РОССИЯ в РОМИП2007: поиск и классификация // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008. 258 с.

229. Агеев М.С., Добров Б.В., Лукашевич Н.В., Штернов С.В. УИС РОССИЯ в РОМИП 2008: поиск и класификация нормативных документов. Российский семинар по Оценке Методов Информационного Поиска // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008. 258 с.

230. Агеев М.С., Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация текстов: методы и проблемы // Ученые записки Казанского государственного университета. Серия Физико-математические науки. 2008. Том 150. книга 4. С. 25-40.

231. Азарова И.В., Синопальникова А.А., Яворская М.В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004. М., 2004. С. 542-547.

232. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика, 1989.

233. Алексеев А.А., Лукашевич Н.В. Автоматическое порождение обновления к аннотации новостного кластера // Труды конференции RCDL-2010. 2010.

234. Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. 2011. N 4. С. 95-103.

235. Антонов А.В., Курзинер Е.С. Автоматическое определение тематики большого необработанного текстового массива // Труды международной конференции Диалог-2002. 2002.

236. Богомолова А.В., Дышкант Н.Ф., Юдина Т.Н. Университетская информационная система РОССИЯ: ресурсы и сервисы для поддержки общественного участия и задач государственного управления // Труды XI Всероссийской объединенной конференции "Интернет и современное общество". Санкт Петербург, 2008. С. 196-199.

237. Большакова Е.И, Большаков И.А., Котляров А.П. Расширенный эксперимент по автоматическому обнаружению и исправлению русских малапропизмов // Труды Международной конференции Диалог-2006. М., 2006. С. 78-83.

238. Воронцов К.В. Машинное обучение. Курс лекций.
http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

239. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. Санкт-Петербург: Изд-во "Питер", 2000. 382 с.

240. Гальперин И.О. Текст как объект лингвистического исследования. М.: Наука, 1981.

241. Герд А.С. Прикладная лингвистика. Изд-во Санкт-Петербургского университета, 2005.

242. Городецкий Б.Ю. Термин как семантический феномен (в контексте переводческой литературы. Код доступа: (http://www.dialog-21.ru/dialog2006/materials/html/GrodetskiyB.htm). 2006.

243. ГОСТ 7.66.-92. Индексирование документов Общие требования к систематизации и предметизации. 1992.

244. ГОСТ 7.74-96. Информационно-поисковые языки. Термины и определения Межгосударственный стандарт 7.74-96. Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 1996.

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

246. ГОСТ 7.59.-2003. Индексирование документов. Общие требования к систематизации и предметизации. Минск: Межгосударственный совет по стандартизации, метрологии и сертификации, 2003.

247. Гринев-Гриневич С.В. Терминоведение. М.: Академия, 2008.

248. ван Дейк Т.А., Кинч В. Стратегии понимания связного текста. // Новое в зарубежной лингвистике. Вып. 23. М.: Прогресс, 1988. С. 153-211.

249. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. М.: Интуит, 2009.

250. Добров Б.В., Лукашевич Н.В. Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск, 2001. С.78-82.

251. Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. М.: Физматлит, 2002. Т.1. С.178-186.

252. Добров Б.В., Лукашевич Н.В. Организация двуязычного поиска в Университетской системе РОССИЯ // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Четвертой Всероссийской научной конференции RCDL-2002. Дубна: ОИЯИ, 2002. Т.2. С.148-158.

253. Добров Б.В., Лукашевич Н.В., Невзорова О.А. Технология разработки онтологий новых предметных областей // Труды Казанской шкрлы по компьютерной лингвистике TEL-2002. Выпуск 7 / Под ред. В.Г.Бухараева, В.Д. Соловьева, Д.Ш.Сулейменова. Казань: Отечество, 2002. С. 90-106.

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

255. Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования практической онтологии // Известия РАН. Теория и системы управления. 2004. N 2. С. 58-68.

256. Добров Б.В., Лукашевич Н.В. Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // Восьмая национальная конференция по искусственному интеллекту КИИ-2002. Коломна, 2002.

257. Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н. Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные

методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL’2005). Ярославль: ЯрГУ им.П.Г.Демидова, 2005. С. 70-79.

258. Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: описания понятий и лексических значений. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог'2005 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П. Селегея. М.: Наука, 2005. С. 138-142.

259. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // Десятая национальная конференция по искусственному интеллекту с международным участием КИИ-2006. М.: Физматлит, 2006. С.489-497.

260. Добров Б.В., Лукашевич Н.В. Вторичное использование лингвистических онтологий: изменение в структуре концептуализации // Восьмая Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». 2006.

261. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Ученые записки Казанского Государственного Университета. Серия Физико-математические науки. 2007. т.149. книга 2. С.49-72.

262. Добров Б.В., Лукашевич Н.В. Транзитивные нетаксономические отношения в онтологическом моделировании. // Труды симпозиума Онтологическое моделирование. Институт проблем информатики РАН, 2008. С. 229-259.

263. Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования практической

онтологии // Известия РАН. Теория и системы управления. 2004. N 2. С. 58-68.

264. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Ученые записки Казанского Государственного Университета. Серия Физико-математические науки. 2007. т. 149. книга 2. С.49-72.

265. EUROVOC. Информационно-поисковый тезаурус. Русская версия тезауруса EUROVOC. Том 1. Алфавитно-пермутационное представление. М.: Издание Государственной Думы, 2001. 500 с.

266. Жинкин Н.И. Механизмы речи. М., 1958.

267. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. Москва: Логос, 2006.

268. Клещев А.С., Шалфеева Е.А. Классификация свойств онтологий. Онтологии и их классификации // НТИ сер. 1. 2005. N 9. С. 16-22.

269. Кронгауз М.А. Семантика. М.: РГГУ, 2001.

270. Леонтьева Н.Н. Семантика связного текста и единицы информационного анализа // НТИ, сер. 2. 1981. N1.

271. Лингвистический энциклопедический словарь / под ред. В.Н. Ярцевой. М.: Советская энциклопедия, 1990.

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

273. Лукашевич Н.В. Разрешение многозначности терминов в процессе автоматического индексирования // Труды международного семинара Диалог'96. Москва, 1996. С.142-146.

274. Лукашевич Н.В., Салий А.Д. Тезаурус для автоматического рубрицирования и индексирования: разработка, структура, ведение // НТИ. Сер.2. 1996. N 1. С.1-6.

275. Лукашевич Н.В., Добров Б.В. Построение и использование тематического представления содержания документов // 5ая Национальная конференция КИИ-96. Казань, 1996. С. 130-134.

276. Лукашевич Н.В. Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2. 1996. N 10. С. 22-30.

277. Лукашевич Н.В. Автоматическое построение аннотаций на основе тематического представления текста // Труды международного семинара Диалог'97. Москва, 1997. С. 188-191.

278. Лукашевич Н.В., Салий А.Д. Представление знаний в системе автоматической обработки текстов // НТИ, Сер.2. 1997. N3. С. 1-6.

279. Лукашевич Н.В., Добров Б.В. Построение структурной тематической аннотации текста // Труды международного семинара Диалог-98. 1998. Том 2. С. 795-802.

280. Лукашевич Н.В. От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Труды международного семинара Диалог-99, Том 2. 1999. С.184-190.

281. Лукашевич Н.В., Добров Б.В. Модификаторы концептуальных отношений в тезаурусе для автоматического индексирования // НТИ, Сер.2. 2001. N 4. С. 21-28.

282. Лукашевич Н.В., Добров Б.В. Исследования тематической структуры текста на основе большого лингвистического ресурса // Труды международного семинара “Диалог 2000”. Том 2. 2000. С. 252-258.

283. Лукашевич Н.В., Добров Б.В. Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // Труды международного семинара Диалог-2001. 2001. С.273-279.

284. Лукашевич Н.В., Добров Б.В. Модификаторы концептуальных отношений в тезаурусе для автоматического индексирования // НТИ, Сер.2. 2001. N 4. С. 21-28.

285. Лукашевич Н.В., Добров Б.В. Автоматическое выявление лексической связности текста // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2001, Выпуск 6. Казань: Отечество, 2001. С.19-38.

286. Лукашевич Н.В., Добров Б.В. Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни. М.: Наука, 2002. Т. 2. С.338-346.

287. Лукашевич Н.В., Добров Б.В. Организация тезаурусного поиска в Университетской информационной системе РОССИЯ // Русский язык в Интернете / Под ред. В.Д.Соловьева. Казань: Отечество, 2003. С. 84-96.

288. Лукашевич Н.В., Добров Б.В. Двуязычный информационный поиск на основе автоматического концептуального индексирования // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог-2003. / Под ред. И.М.Кобозевой, Н.И.Лауфер, В.П.Селегея. М.: Наука, 2003. С.425-432.

289. Лукашевич Н.В., Добров Б.В. Разграничение общезначимой лексики и терминологии и автоматическая обработка больших электронных коллекций // Русский язык: исторические судьбы и современность. II Межд. конгресс исследователей русского языка. М.: МГУ, 2004. С. 481-482.

290. Лукашевич Н.В, Добров Б.В. Отношения в онтологиях для решения задач информационного поиска в больших разнородных текстовых коллекциях // Девятая национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды коференции. В 3-х т. М.: Физматлит, 2004. Т2. С.544-551.

291. Лукашевич Н.В., Добров Б.В. Разрешение лексической многозначности на основе тезауруса предметной области. Компьютерная лингвистика и интеллектуальные технологии. // Труды международной конференции «Диалог 2007». М.: Наука, 2007. С. 400-406.

292. Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостного кластера на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии по материалам ежегодной Международной конференции «Диалог 2009». 2009. Выпуск 8 (15), С. 299-305.

293. Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний // Интернет-математика 2007: Сборник работ участников конкурса. Екатеринбург: Изд-во Урал. ун-та, 2007. С.108-117.

294. Лукашевич Н.В. Моделирование отношения ЧАСТЬ-ЦЕЛОЕ в лингвистических и онтологических ресурсах // Информационные технологии. 2007. N 12. С. 28-34.

295. Лукашевич Н.В. Проблемы установления родовидовых отношений в лингвистических онтологиях // Материалы Всероссийской конференции «Знания-Онтологии-решения» (ЗОНТ-07). 2007. С. 211-220.

296. Лукашевич Н.В. Типы и роли в лингвистических онтологиях // Труды Казанской школы по компьютерной лингвистике TEL-2006. Казань: Отечество, 2007. С.49-64.

297. Лукашевич Н.В. Квазисинонимы в лингвистических онтологиях. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог 2010». 2010. С. 307-312.

298. Лукашевич Н.В., Логачев Ю.М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование, разд. 2. 2010. С. 108-116.

299. Лукашевич Н.В. Понятия в формальных и лингвистических онтологиях // Научно-техническая информация, сер.2. 2011. N 7. С. 1-8.

300. Мальковский М.Г., Соловьев С.Ю. Универсальное терминологическое пространство // Труды международного семинара

"Компьютерная лингвистика и интеллектуальные технологии". М: Наука, 2002. т.1. С.266-270.

301. Некрестьянов И., Некрестьянова М. Особенности организации и проведения РОМИП 2008. Код доступа: http://romip.ru/romip2008/2008_01_organizers.pdf

302. Никитина С.Е. Семантический анализ языка науки. М.: Наука, 1987.

303. Новиков А.И. Семантика текста и ее формализация. М.: Наука, 1983.

304. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М.: Физматлит, 1997.

305. Севбо И.П. Структура связного текста и автоматизация реферирования. М.: Изд-во Наука, 1969.

306. Сегалович И., Маслов М. Яндекс на РОМИП-2004. Некоторые аспекты полнотекстового поиска и ранжирования Яндекса // РОМИП-2004, 2004.

307. Суперанская А.В., Подольская Н.В., Васильева Н.В. Общая терминология: Вопросы теории / Отв. Ред. Т.Л.Канделаки. Изд. 2-е, стереотипное. М.: Едиториал УРСС, 2003.– 248 с.

308. Сухоногов А.М., Яблонский С.А. Автоматизация построения англо-русского WordNet. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2005 / Под ред. А.С.Нариньяни. М.: Наука, 2005.

309. Тузовский А.Ф., Чириков С.В., Ямпольский В.З. Системы управления знаниями (методы и технологии). Томск: Изд-во научно-технической литературы, 2005.

310. Уемов А.И. Вещи, свойства и отношения. М., 1963.

311. УКАЗ Президента РФ от 15.03.2000 N 511. О Классификаторе правовых актов.

312. Хорошевский В.Ф. Онтологические модели и Semantic Web: откуда и куда мы идем? // Труды семинара «Онтологическое моделирование» под редакцией Калиниченко Л.А. Москва: ИППИ РАН, 2008. С. 13-45.

313. Шевченко Н.В. Основы лингвистики текста. М.: Приор-издат, 2003.

314. Шелов С.Д. Определение терминов и понятийная структура терминологии. Изд-во С.-Петербургского Университета, 1998.

315. Шелов С.Д. Термин. Терминологичность. Терминологические определения. СПб., 2003. 280 с.

316. Шемакин Ю.И. Тезаурус в автоматизированных системах управления и информации. М.: Военное изд-во министерства обороны СССР, 1974. 192 с.

317. Шемакин Ю.И. Тезаурус научно-технических терминов. М.: Военное изд-во министерства обороны СССР, 1974.