Знания-Онтологии-Теории (ЗОНТ-09)

Тезаурус РуТез как ресурс для решения задач информационного поиска

Б.В. Добров[1,2], Н.В. Лукашевич[1,2]

[1]*Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова, Ленинские горы, д.1, стр. 4, 119992, Москва, Россия*
[2]*АНО Центр информационных исследований, Ленинские горы, д.1, стр. 77, 119992, Москва, Россия*

dobroff,louk@mail.cir.ru

Аннотация. *Описываются принципы построения информационно-поискового тезауруса РуТез, специально разрабатываемого в течение 15 лет для решений задач автоматической обработки текстов. Наш опыт показывает, что структура тезауруса должна быть специально адаптирована к задачам информационного поиска, а тезаурусные технологии не должны противопоставляться современным технологиям пословной обработки текстов, а органично учитывать последние достижения в этой сфере. При учете таких условий применение тезаурусов может дать улучшение качества решения задачи по сравнению с лучшими пословными методами.*
Ключевые слова: информационно-поисковые тезаурусы, информационный поиск, автоматическая обработка текстов, автоматическая рубрикация

1 Введение

Область современного информационного поиска чрезвычайно разнообразна. Она включает 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

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.

Абсолютно подавляющее число методов информационного поиска решает эти задачи на основе минимальных дополнительных предварительных знаний и базируется на моделях текста как набора слов (“bag of words”), предлагая изощренные методы учета частотностей встречаемости слов в предложении, тексте, наборе документов, совместной встречаемости слов и т.п.

Такое представление текстов как простого набора слов имеет большое количество очевидных недостатков, затрудняющих поиск релевантных текстов, таких как

- избыточность -- в пословном индексе используются слова-синонимы, выражающие одни и те же понятия;
- предположение о независимости слов текста - слова текста считаются независимыми друг от друга, что не соответствует свойствам связного текста;
- многозначность слов -- поскольку многозначные слова могут рассматриваться как дизъюнкция двух или более понятий, выражающих различные значения многозначного слова, то маловероятно что все элементы этой дизъюнкции интересуют пользователя.

Потенциально эти проблемы могли бы решаться на базе лингвистических и/или онтологических ресурсов (онтологий, тезаурусов) в рамках технологий так называемого концептуального индексирования. Однако на практике тезаурусы и онтологии редко применяются в промышленных информационно-поисковых системах, основанных на автоматической обработке текстов. Такая ситуация связана с целым рядом обстоятельств.

Во-первых, если предлагается использовать некоторый лингвистический ресурс, то он должен включать описания десятков тысяч слов и словосочетаний. Процент ошибок ресурса должен настолько мал, чтобы не испортить возможные улучшения, получаемые от применения этого ресурса. При этом нужно понимать, что ведение любого лингвистического ресурса всегда будет отставать от развития предметной области, то есть даже наиболее качественный лингвистический ресурс будет всегда неполон.

Во-вторых, применение тезаурусов и онтологий требует высокого качества разрешения многозначности слов. Так, в работе [7] 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 исследовании [2] авторы ставят эксперимент по измерению чувствительности качества информационного поиска к ошибкам разрешения многозначности. В данной работе используются документы, которые вручную размечены по значениям WordNet [6]. Для эксперимента в разметку по значениям искусственно вносится некоторый процент ошибок.

Внесение ошибок разрешения многозначности в индексирование по синсетам WordNet показало, что 10% ошибок не влияет на качество поиска, что находится в соответствии с работой [7]. При этом выяснилось, что при уровне 30% ошибок качество поиска на основе концептуального индекса по синсетам WordNet превосходит поиск по классической векторной модели, основанной на пословном индексе [1]. Таким образом, авторы работы [2] делают вывод, что если выполнять разрешение многозначности с точностью больше 70%, то это даст преимущество по сравнении с пословными векторными моделями. Однако лучшие методы разрешения многозначности на основе тезауруса WordNet дают не более 62-65% точности разрешения многозначности на множестве слов текста [5, 8].

В-третьих, применение отношений тезауруса или онтологии для расширения запросов может столкнуться с проблемой неточно описанных отношений или отношений, которые не соответствуют контексту запроса. Поэтому такое применение отношений часто ведет к возрастанию полноты поиска и одновременно к значительному снижению точности поиска. Так, в последнее время глобальные поисковые системы Yandex и Google стали активно применять расширение запросов однокоренными словами, что может рассматриваться как минимальный тезаурус, но при многих запросах даже такое минимальное расширение запроса может оказаться нерелевантным.

Наконец, существует мнение, что применяемые статистические методы имплицитно учитывают лингвистическую информацию, то есть то, что лингвистические методы учитывают эксплицитно. Также известно мнение, что текст – это лишь набор характеристик (features), которые хорошо учитываются статистическими моделями. В качестве примеров моделирования лингвистических подходов статистическими методами Хелен Ворхес [9] приводит следующие примеры: морфологический анализ может быть приближен стеммингом, извлечение словосочетаний - выявлением часто встречающихся пар слов, процедуры разрешения многозначности могут смоделированы мерами сходства контекстов.

В течение уже почти 15 лет мы разрабатываем тезаурус РуТез, предназначенный для применения в автоматических режимах обработки текста для решения задач информационного поиска [15, 16], и проводим эксперименты по применению тезауруса в различных приложениях информационного поиска таких как автоматическое концептуальное индексирование, расширение запросов, автоматическая рубрикация текстов, автоматическое аннотирование документов, автоматическая кластеризация [11, 12, 14, 17, 18].

Тезаурус РуТез является поисковым инструментом Университетской информационной системы Россия [20], содержащей более 2 миллионов документов жанра деловой прозы – нормативные акты, материалы общественно-политических СМИ и т.д.

Были сделаны серьезные усилия, чтобы разобраться, какая структура тезауруса является наиболее подходящей для применения в задачах информационного поиска в широких предметных областях, требующих описания в тезаурусе десятков тысяч слов и словосочетаний. Также задачей наших

исследований было выяснить, какое улучшение качества обработки текстов по сравнению с пословными моделями и при каких условиях может быть получено.

Данная статья посвящена описанию принципов разработки тезауруса РуТез. Также будут представлены результаты экспериментов с тезаурусом при решении различных задач.

2 Принципы разработки тезауруса РуТез

2.1 Общая структура тезауруса

Тезаурус – это иерархическая сеть понятий. Каждое понятие имеет имя. Для сопоставления с текстом каждое понятие снабжается набором текстовых выражений (=текстовых входов, =терминов), значения которых соответствуют данному понятию. В качестве таких текстовых входов могут выступать однословные существительные, прилагательные, глаголы, именные и глагольные группы. Количество таких текстовых входов понятий может быть достаточно велико, например, превышать 20 единиц. При вводе нового понятия делаются специальные усилия, чтобы максимально подробно перечислить его возможные текстовые входы. Каждое понятие связывается отношениями с другими понятиями Тезауруса.

В настоящее время Тезаурус РуТез содержит более 51.5 тысяч понятий, более 155 тысяч текстовых входов (слов, словосочетаний), более 200 тысяч отношений между понятиями, с учетом иерархии тезаурусных связей всего устанавливается более 2 миллионов отношений между понятиями. Понятия тезауруса снабжены также английскими текстовыми входами (более 125 тысяч слов и словосочетаний).

В составе тезауруса РуТез выделяется так называемый Общественно-политический тезаурус, содержащий лексику и терминологию, относящуюся к широкой области современной общественной жизни общества, включая такие сферы как экономика, политика, социальные вопросы, спорт, искусство и др. В настоящее время Общественно-политический тезаурус включает порядка 35 тысяч понятий и около 100 тысяч текстовых входов. Понятия тезауруса, не входящие в состав Общественно-политического тезауруса, составляют так называемый Общий лексикон. Понятия и языковые выражений, отнесенные к Общему Лексикону, могут встретиться в разных предметных областях, выражают независимые от предметной области сущности, свойства, отношения, эмоции, оценки. При разметке предметными областями тезауруса WordNet авторы работы [4] также выделили область понятий, сходную с нашим Общим Лексиконом, и назвали ее Factotum.

В подавляющем большинстве приложений мы используем Общественно-политический тезаурус, к которому может быть добавлена совокупность понятий из Общего лексикона, необходимых для обработки текстов в данной предметной области. Это связано, прежде всего, с тем, что Общественно-политический тезаурус содержит значительно меньший процент многозначных языковых выражений. Разные значения Общественно-политического тезауруса либо относятся к разным предметным областям и поэтому неплохо отделяются друг от друга, либо, наоборот, соответствующие понятия тесно связаны друг с другом, и, таким образом, их возможное неразличение может и не приводить к серьезным последствиям в приложениях информационного поиска [3].

В работе [18] мы показали, что средняя точность разрешения многозначности для текстовых выражений описанных в Общественно-политическом тезауруса по разным источникам (газеты, новостные сообщения) составила около 73%, что превышает уровень точности разрешения многозначности (70%), который указан в работе [2] в качестве уровня, позволяющего получать преимущество по сравнению с пословными моделями представления содержания текста.

В некоторых задачах, например, в задаче автоматической рубрикации текстов вся обработка основывается на концептуальном индексе, построенном на базе понятий Общественно-политического тезауруса [14]. В других задачах (например, в информационном поиске, кластеризации текстов) используется смешанный индекс как по словам, так и по понятиям Общественно-политического тезауруса [11], и тогда для слов из Общего лексикона (то есть не входящих в состав Общественно-политического тезауруса) разрешение многозначности не производится.

2.2 Принципы описания отношений в тезаурусе РуТез

В результате исследований и экспериментов мы пришли к выводу, какими должен быть набор отношений ресурса, предназначенного для автоматической эффективной работы в информационно-поисковых приложениях. Основным принципом для описания отношений в тезаурусе,

предназначенном для использования в автоматических режимах обработки текстов, является следующий [17]:

> *Среди потенциального множества отношений понятия наиболее стабильно можно опираться на те отношения, которые не исчезают, не меняются в течение всего срока существования любого или подавляющего большинства экземпляров понятия.*

Например, любой лес всегда состоит из деревьев.

В настоящее время, в тезаурусе имеется четыре основных типа отношений, которые мы относим к такому множеству надежных отношений.

Первый тип отношений – родовидовое отношение НИЖЕ-ВЫШЕ, обладает свойством транзитивности и наследования.

Второе тип отношений – отношение ЧАСТЬ-ЦЕЛОЕ. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого, и не относиться к чему-либо другому.

В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения ЧАСТЬ-ЦЕЛОЕ, что очень важно для автоматического вывода в процессе автоматической обработки текстов [19].

Еще один тип отношения, называемого несимметричной ассоциацией АСЦ2-АСЦ1, связывает два понятия, которые не могут быть связаны рассмотренными выше отношениями, но когда одно из которых не существовало бы без существования другого. Например, понятие *САММИТ* требует существования понятия *ГЛАВА ГОСУДАРСТВА*.

Последний тип отношений – симметричная ассоциация связывает понятия очень близкие по смыслу, но которые разработчики не решились склеить в одно понятие.

Отношения ВЫШЕ-НИЖЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии.

3 Эксперименты с использованием тезауруса РуТез

При применении тезауруса РуТез в задачах информационного поиска предполагается, что строится концептуальных индекс по Общественно-политическому тезаурусу. Построение такого индекса включает в себя обработку текста графематическим и морфологическим анализаторами, сопоставление текста с тезаурусом, разрешение многозначности слов и выражений.

На основе выделенных в тексте понятий и их отношений между собой строится тематическое представление текста, в процессе которого все понятия текста разбиваются на совокупности близких по смыслу понятий - так называемые тематические узлы, которые состоят из центра тематического узла и понятий, близких по тезаурусной окрестности к тематическому центру.

Тематические узлы разделяются на главные и локальные, что соответствует глобальной теме текста и локальным подтемам [16]. При построении тематического представления делается попытка учесть отношения между понятиями тезауруса, в частности, для того, чтобы относительно малочастотные понятия текста могли получить более высокие веса, если в тексте имеется достаточное количество понятий, близких им по смыслу.

Место понятия тезауруса в тематическом представлении определяет оценку значимости $\omega(d; D)$ понятия относительно содержания текста – наибольший вес получают центры главных тематических узлов, наименьший – понятия, не входящие в главные или локальные тематические узлы. Окончательный вес понятия на основе тематического представления сглаживается с учетом частотности понятия в документе [12, 14]. Полученный вес понятия используется в концептуальном индексе документа.

В данном разделе будут рассмотрены три эксперимента, в которых используется Общественно-политический тезаурус и построенный по нему концептуальный индекс, и будут приведены результаты оценок качества решения задач.

3.1 Эксперимент по тестированию качества поиска по Общественно-политическому тезаурусу на простых запросах

В данном разделе мы опишем эксперимент по оценке качества поиска с использованием тезаурусных знаний в условиях, когда задаваемые запросы хорошо покрываются текстовыми входами Общественно-политического тезауруса.

Запросы в информационной системе могут состоять из различного числа терминов и слов. С точки зрения тезауруса, простейшим запросом является запрос, ссылающийся на одно понятие тезауруса. Все другие запросы, ссылающиеся на два или более понятий, должны обрабатываться как функция от элементарного запроса.

Мы предполагаем, что потенциальное качество расширения запроса на базе отношений тезауруса может изучаться на простых запросах. Если поисковые характеристики расширения элементарных запросов являются низкими, то качество расширения сложных поисковых запросов не может быть лучше. Если тезаурусные отношения дают возможность эффективного расширения запроса для простых случаев, то это является важным шагом для изучения способов расширения сложных запросов.

В работе [16] описывается эксперимент по оценке поиска с расширением по тезаурусу для простых запросов. Для эксперимента мы исполнили набор запросов в УИС РОССИЯ [20]. Каждый запрос был сформулирован дважды: один раз как запрос на поиск по словам, второй раз - как запрос на поиск по понятиям тезауруса с расширением по дереву, то есть запрос расширялся на все понятия тезауруса, которые описаны как иерархически нижестоящие к исходному понятию. В качестве запросов были выбраны рубрики из Классификатора правовых актов (одобрен Указом Президента РФ от 15 марта 2000 г.). Поиск осуществлялся на 40-тысячной коллекции нормативных актов УИС РОССИЯ.

При выполнении подавляющего количества запросов количество документов, найденных с использованием деревьев Тезауруса, значительно превышало количество документов, найденных по словам. Полнота поиска с использованием деревьев тезауруса значительно возросла. Однако, как известно, увеличение полноты поиска часто сопровождается снижением точности поиска, то есть 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 [9]. Точность выполнения запроса вычисляется при следующих трех значениях полноты: 0.2, 0.5, 0.8.

Всего было выполнено тестирование 19 запросов – рубрик Классификатора правовых актов. Средняя точность по трем точкам по тезаурусу составила – 0.62, по словам – 0.44.

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.

3.2 Тезаурус и векторная модель в задаче поиска по коллекции нормативно-правовых актов

В реальных условиях задания запросов пользователем запросы по отношению к тезаурусу 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 термины тезауруса,

- 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 концептов на основе тезауруса, нахождение сущностей из запроса в ограниченном числе предложений документа. Модель тестировалась на семинаре РОМИП-2008 в коллекции нормативно-правовых документов [11].

Основной направленностью разработки модели была обработка длинных информационных запросов, то есть 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 концептов Общественно-политического тезауруса [2].

Концепты 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, то есть каждый документ получает вес по следующей формуле:

$$W_d = \alpha_1 W_{word} + (1- \alpha_1) W_{conc,}$$

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 концептов тезауруса в предложениях.

Наконец, **на третьем этапе** исходный вес документа**,** полученный на первом этапе, замешивается с весом документа по предложениям, полученный на втором этапе.

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 асессорами, показатель средней точности MAP [13]- 0.296 (см. Рис.1), который превышает показатель следующего участника – 0.276 на 7%.

Чтобы проанализировать, насколько хорошо модель отработала на целевом множестве длинных информационных запросов, мы разбили запросы на несколько групп, отдельно выделив длинные информационные запросы, длиной более 3 слов, например, *уплата налога на прибыль организацией при отсутствии затрат* (всего 27 запросов из 95 оцениваемых).

Пользуясь этой классификацией мы разделили все оцененные запросы этой дорожки на соответствующие группы и оценили среднюю точность участников по этим группам. На длинных информационных запросов нами была получена средняя точность MAP – 0.36, что значительно превышает наш средний результат (0.29), а также результат следующего участника (0.32).

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.

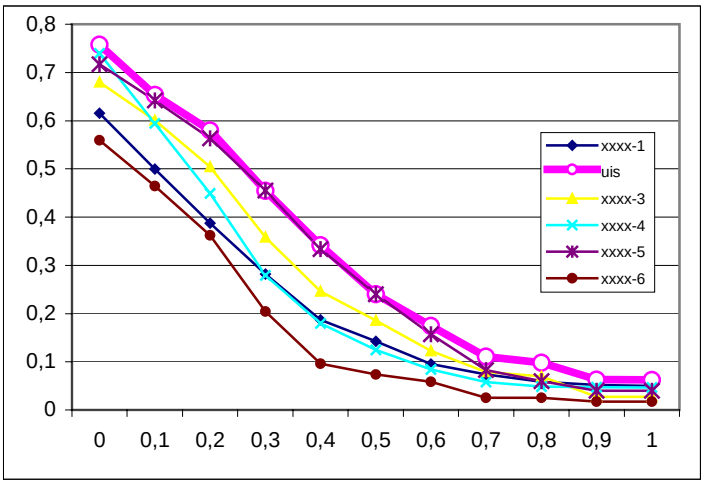

Рис.1. Результаты дорожки РОМИП-2008 Legal adhoc (pd35).

3.3 Тезаурус РуТез в задаче автоматической рубрикации текстов

В данном разделе мы опишем эксперимент по применению Общественно-политического тезауруса в задаче автоматической рубрикации текстов.

Автоматическая рубрикация является одной из классических задач информационного поискав. При решении этой задачи существует определенное противостояние между методами машинного обучения (когда по имеющемуся обучающему множеству автоматически строится решающее правило) и, так называемым, «инженерным подходом» (когда решающее правило формируется экспертами).

Традиционным нашим подходом в сфере автоматической рубрикации является инженерный подход, в котором содержание рубрики описывается как булевское выражение над понятиями Общественно-политического тезауруса. В рамках этого подхода было построено около 15 систем автоматической рубрикации текстов с размерами рубрикаторов от нескольких десятков рубрик до трех тысяч рубрик. Построенные системы автоматической рубрикации были предназначены для обработки разных типов документов таких как нормативно-правовые документы, научные публикации в области гуманитарных наук, новостные сообщения, социологические опросы [14].

На семинаре РОМИП 2007 в рамках решения задачи по автоматической классификации документов коллекции ROMIP.BY по 247 рубрикам коллекции ROMIP.dmoz – у нас была возможность сравнить два этих подхода. Одной из мотиваций эксперимента было получить оценки трудоемкости для построения описания рубрикатора [12].

Недостатком использования «инженерного подхода» для решения задачи автоматической рубрикации считается высокая трудоемкость описания решающего правила рубрики. Мы используем «инженерный подход» в течение достаточно длительного времени (с 1996 года) и за это время разработали технологию, снижающую трудоемкость работы. Основная идея нашего подхода к автоматической рубрикации текста состоит в существенном использовании иерархии тезаурусных связей. То есть смысл рубрики описывается достаточно короткой формулой над понятиями тезауруса, а затем производится автоматическое расширение построенной формулы.

Подробнее – смысл рубрики описывается как дизъюнкция:

$$R = \bigcup_i D_i$$

В данном эксперименте всего было описано 234 рубрики из 247. Для 234 рубрик описано 265 дизъюнктов.

Каждый дизъюнкт представляется конъюнкцией (всего 334 конъюнкта).

$$D_i = \bigcap_j K_{ij}$$

Каждый конъюнкт представляется в виде совокупности «положительных» и «отрицательных» «опорных концептов», которые задают в регулируемом порядке применения правила добавления и удаления множеств подчиненных по иерархии концептов (здесь *f(.)* – схематическое обозначения правил учета иерархии):

$$K_{ij} = \bigcup_m f_m(c_{ijm}) \setminus \bigcup_n f_n(e_{ijn})$$

Всего было использовано 899 опорных концептов.

$$R = \bigcup_i D_i = \bigcup_i \left[\bigcap_j K_{ij} \right] = \bigcup_i \left[\bigcap_j \left(\bigcup_k d_{ijk} \right) \right]$$

Далее, расширяя по иерархии тезауруса, получаем полное представления рубрики (где уже для всех рубрик задействовано 40161 концептов - естественно, с учетом повторения – и 107897 текстовых входов).

Вес рубрики определяется как максимум весов дизъюнктов. Вес дизъюнкта определяется как взвешенная сумма весов конъюнктов с учетом совместной встречаемости конъюнктов в тексте [12, 14].

При описании рубрик классификатора эксперты, в основном, ориентировались на формулировку рубрики. В единичных случаях эксперты заходили на сайт dmoz.org для уточнения объема рубрики. Всего для решения задачи описания рубрикатора ROMIP.dmoz было затрачено 8 человеко-часов двух экспертов (2 эксперта по 4 часа).

Общественно-политический тезаурус покрывает практически все предметные области, отражаемые в деловой прозе – нормативных актах, СМИ федерального уровня. Поэтому для решения задачи описания рубрик потребовалось ввести в тезаурус дополнительно только восемь понятий для описания специфических экстремальных видов спорта.

Рассмотрим пример описания на примере рубрики №43 «домашний ремонт».

```
(     РЕМОНТ(N)              OR  КАПИТАЛЬНЫЙ РЕМОНТ(N)
 OR   ТЕКУЩИЙ РЕМОНТ(N)      OR  РЕМОНТНО-СТРОИТЕЛЬНЫЕ РАБОТЫ(N) )
AND
(     ЖИЛОЕ ЗДАНИЕ(L) OR  ЖИЛОЕ ПОМЕЩЕНИЕ(L) OR  КВАРТИРА(L) )
```

здесь пометка «L» означает, что предусматривается только расширение по отношениям типа «ВЫШЕ-НИЖЕ», пометка «N» означает отсутствие расширения.

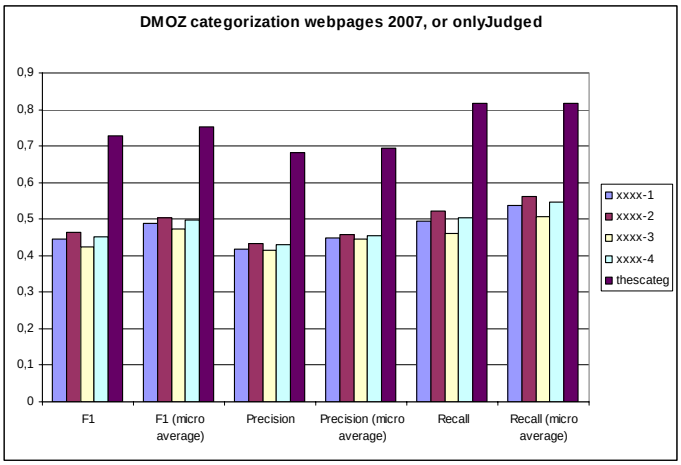

Рис.2. РОМИП-2007 классификация веб-страниц, (or)

На Рис.2 приведены результаты для дорожки классификации Веб-страниц коллекции ROMIP.BY. Достижение показателей качества рубрицирования [13] 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 с серьезной противоречивостью коллекции, а именно, со следующими обстоятельствами. Как база для обучения были представлены данные ручной рубрикации сайтов «как целого». Однако внутри сайта могут быть представлены достаточно разные по содержанию страницы, что и затрудняет выработку разделяющих правил методами машинного обучения.

Полученные результаты позволяют сделать следующие выводы.

Существуют задачи классификации текстов, когда нет достаточно качественной обучающей коллекции, например, нет достаточного множества обучающих примеров или ручная классификация проведена недостаточно последовательно. В таких условиях применения методов машинного обучения очень проблематично.

При машинном обучении системы извлекают некоторые знания о языке и мире, которые можно условно подразделить на: общие знания о языке и мире, необходимые для работы различных приложений в разнообразном круге предметных областей, и текущие знания, характерные именно для текущей задачи, текущей коллекции, данного типа пользователей и т.п.. Значимую часть общих знаний о современной жизни общества и современном языке деловой прозы нам удалось упорядочить в рамках понятийных структур Общественно-политического тезауруса.

В описываемом эксперименте у нас не было возможности сделать предварительный прогон, оценить и исправить ошибки и неточности описания рубрик. В обычной практике проводится несколько итераций, консультаций с экспертами. Поэтому имеются определенные возможности улучшения результатов рубрикации, полученных на основе Общественно-политического тезауруса.

Отметим, что, на наш взгляд, определенный потенциал исследований представляет собой исследование возможностей концептуального индекса, построенного на понятиях тезауруса, как признаковой базы для методов машинного обучения. Так, на семинаре РОМИП 2004 наилучший результат в задании автоматической рубрикации нормативных документов был получен методом машинного обучения, формирующим логические формулы на основе понятий Общественно-политического тезауруса [10].

4 Заключение

В течение около 15 лет мы разрабатываем тезаурусы и исследуем технологии применения их для решения различных задач информационного поиска. Созданы тезаурус русского языка РуТез и Общественно-политический тезаурус, содержащий лексику и терминологию в широкой сфере общественных отношений.

В настоящее время наши выводы по применению тезауруса РуТез в информационном поиске таковы:

- структура тезауруса, принципы описания его единиц и отношений имеет важное значение для повышения качества решения задач информационного поиска. Тезаурус РуТез отличается по ряду важных принципов разработки как от тезаурусов типа WordNet, так и от традиционных информационно-поисковых тезаурусов,
- для ряда задач применение тезауруса РуТез, основанное только на сделанных в нем описаниях, может оказаться не лучше применения пословных моделей (из-за возможных проблем нехватки информации в тезаурусе, неточности описаний, проблем разрешения многозначности и др.). Однако гибкое сочетание качественной пословной модели и знаний, описанных в РуТез, дает улучшение качества на 10-15 процентов,
- для ряд задач знание, накопленное в тезаурусе РуТез, позволяет получить решение задачи более быстро и более качественно, чем использование статистических методов и методов машинного обучения.

Литература

[1] Buckley C., Allan J., Salton J. (1993) Automatic Routing and Ad-hoc Retrieval Using Smart: TREC 2. In *Proceedings of the Second Text Retrieval Conference*. NIST Special Publication 500-215, pp.45-56.

[2] 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*.

[3] Gonzalo J. Chugur I., Verdejo F. Sense clustering for information retrieval: evidence from Semcor and the EWN Interlingual Index. In: *Proceedings of the ACL 2000 Workshop on Word Senses and Multilinguality,* 2000.

[4] 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, 2002.

[5] Mihalcea R., Chklovsky T., Kilgarriff A. Framework and results for English SENSEVAL. In: *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, July 2004, Barcelona, Spain. 2004. P.25–28.

[6] Miller G., Beckwith R., Fellbaum C., Gross D. and Miller K., Five papers on WordNet. - *CSL Report* 43. Cognitive Science Laboratory, Princeton University, 1990

[7] Sanderson M. (1994). Word Sense Disambiguation and information retrieval. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[8] 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

[9] Voorhees, E. (1999). Natural Language Processing and Information Retrieval. In: M.T.Pazienza (ed.). *Information Extraction: Towards Scalable, Adaptable Systems*, New York: Springer, pp. 32-48.

[10] Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // *Российский семинар по оценке методов информационного поиска. Труды второго российского семинара РОМИП'2004 (Пущино, 01.10.2004)* – СПб: НИИ Химии СПбГУ. – 2004. – С.62-89.

[11] М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, С.В. Штернов. УИС РОССИЯ в РОМИП 2008: поиск и класификация нормативных документов. // *Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.)* Санкт-Петербург: НУ ЦСИ, 2008, 258 с.

[12] М. Агеев, Б. Добров, П. Красильников, Н. Лукашевич, А. Павлов, А. Сидоров, С. Штернов. УИС РОССИЯ в РОМИП2007: поиск и классификация. // *Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. (Дубна, 9 октября 2008г.)* Санкт-Петербург: НУ ЦСИ, 2008, 258 с.

[13] Агеев М., Кураленок И., Некрестьянов И. Официальные метрики РОМИП'2007. // *Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008*. Санкт-Петербург: НУ ЦСИ, 2008, 258 с. стр.237-247.

[14] Добров Б.В., Лукашевич Н.В., Автоматическая рубрикация полнотекстовых документов по классификаторам сложной структуры // *Восьмая национальная конференция по искусственному интеллекту. КИИ-2002. 7-12 октября 2002*, Коломна – М.: Физматлит – Т.1 – С.178-186.

[15] Лукашевич Н.В., Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // *НТИ. Сер.2.* - 1995. - N 3. - С.21-24.

[16] Лукашевич Н.В., Добров Б.В., Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса // *Труды международного семинара Диалог-2001.* - Аксаково-2001.- с.273-279.

[17] 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 (28 сентября –2 октября 2004 г., Тверь) : Труды коференции. В 3-х т.* - Т2. – М.: Физматлит, 2004. – С.544-551.

[18] 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.Стр.108-117.

[19] Лукашевич Н.В. Моделирование отношения ЧАСТЬ-ЦЕЛОЕ в лингвистическом ресурсе для информационно-поисковых приложений. // *Информационные технологии*, N12, - 2007.

[20] Университетская информационная система Россия (УИС РОССИЯ), http://uisrussia.msu.ru; http://www.cir.ru.