ОПЫТ СОЗДАНИЯ ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ПОИСКОВЫХ МАШИН

THE DEVELOPMENT OF LANGUAGE PROCESSING FOR SEARCH ENGINES: EXPERIENCE AND APPROACH

Чубинидзе К.А. *(konstantinch@convera.ru),* ***Ежов А.С.*** *(AlexanderE@convera.ru),*
Громов А.И. *(alexanderg@convera.ru),* ***Кусова А.Т.*** *(alana@convera.ru)*
ООО CONVERA

В докладе сформулированы современные требования к лингвистическому обеспечению поисковых машин. Дано описание поисковых словарей и алгоритмов системы RetrievalWare компании Convera. Обобщен опыт, полученный при разработке лингвистического процессора русского языка данной системы.

1. Требования к современным поисковым системам

Лингвистическое обеспечение поисковых машин всегда имело отличия от аналогичных ресурсов, предназначенных для извлечения из текстов значимой информации или для машинного перевода. В современных условиях при значительном увеличении области поиска эти отличия стали более заметными.

Лингвистическое обеспечение поисковых машин можно разделить на 3 класса:

1. общее лингвистическое обеспечение, к которому, как правило, относятся модуль морфологического анализа и семантические словари;

2. специальные алгоритмы настройки поискового запроса, предназначенные для увеличения точности поиска;

3. специальные алгоритмы ранжирования и структурирования результатов поиска, предназначенные для сокращения времени нахождения искомого документа среди найденных поисковой машиной.

На первый взгляд, морфологический анализ и семантические словари предназначены для увеличения полноты поиска. Данное утверждение требует пересмотра. Большинство поисковых систем предназначено для пользователей, квалификация которых не позволяет им самостоятельно сформулировать достаточно точный поисковый запрос. Кроме того, в условиях недостаточно определенного направления поиска это не всегда можно сделать. По этим причинам поисковый запрос, как правило, состоит из двух-трех поисковых терминов. Очевидно, что при этом в достаточно представительной области поиска будет найдено больше формально релевантных документов, чем пользователь сможет проанализировать. При этом пользователь, как правило, ограничивается просмотром нескольких документов, получивших наибольшую оценку релевантности. Статистика такова: если список найденных документов содержит более двух экранных страниц, то 85% пользователей просматривают только первые 2-3 документа, 10% - все документы, перечисленные на первой странице списка, 3% заглядывают на вторую страницу списка, 2% просматривают 2 первых страницы списка и иногда заглядывают на последнюю. Лишь некоторые просматривают первую и последнюю страницы для определения разброса оценок совпадений

В связи с этим общее лингвистическое обеспечение должно быть таким, чтобы во взаимодействии с алгоритмами настройки поискового запроса, ранжирования и структурирования результатов его выполнения обеспечивать максимально возможную точность поиска. Наш опыт в этой области мало отличается от мирового и в данном докладе мы попробуем его обобщить.

2. Специфика лингвистического обеспечения поисковой системы RetrievalWare

В системе RetrievalWare применяются три типа семантических словарей, структура которых определяется не только требованиями оптимизации полноты и точности поиска, но и возможностью их применения для динамической классификации документов. Их принципиальное отличие от традиционных словарей синонимов и семантических сетей заключается в точной настройке, благодаря которой для индексации одного документа могут одновременно эффективно использоваться десятки словарей. Каждый словарь содержит набор семантических отношений, которые характеризуются весовыми коэффициентами, связывающими лексические единицы и обозначаемые ими понятия (см. Рисунки 1 и 2). Отношения имеют направленный характер от понятия к лекси-

ческой единице, например мероним, синоним, антоним, ассоциации. Пользователи могут определить дополнительные отношения, описывающие семантику предметной области словаря.

Понятия могут быть связаны в иерархии таксономическими отношениями в соответствии с различными принципами деления предметной области, например, часть-целое или род-вид. Для динамической классификации могут применяться десятки классификаторов, которые представляют собой иерархии условий, состоящих из понятий, и, при необходимости, связывающих их логических выражений. Семантические словари и таксономии позволяют моделировать непересекающиеся предметные области, а классификаторы – интересующие пользователя проблемы, которые могут находиться на пересечении предметных областей. Таким образом, точность поиска обеспечивается семантической близостью терминов-расширений запроса и их принадлежностью к предметной области поиска, а скорость обработки результатов поиска – быстрой навигацией по списку найденных документов с помощью одного или двух классификаторов.

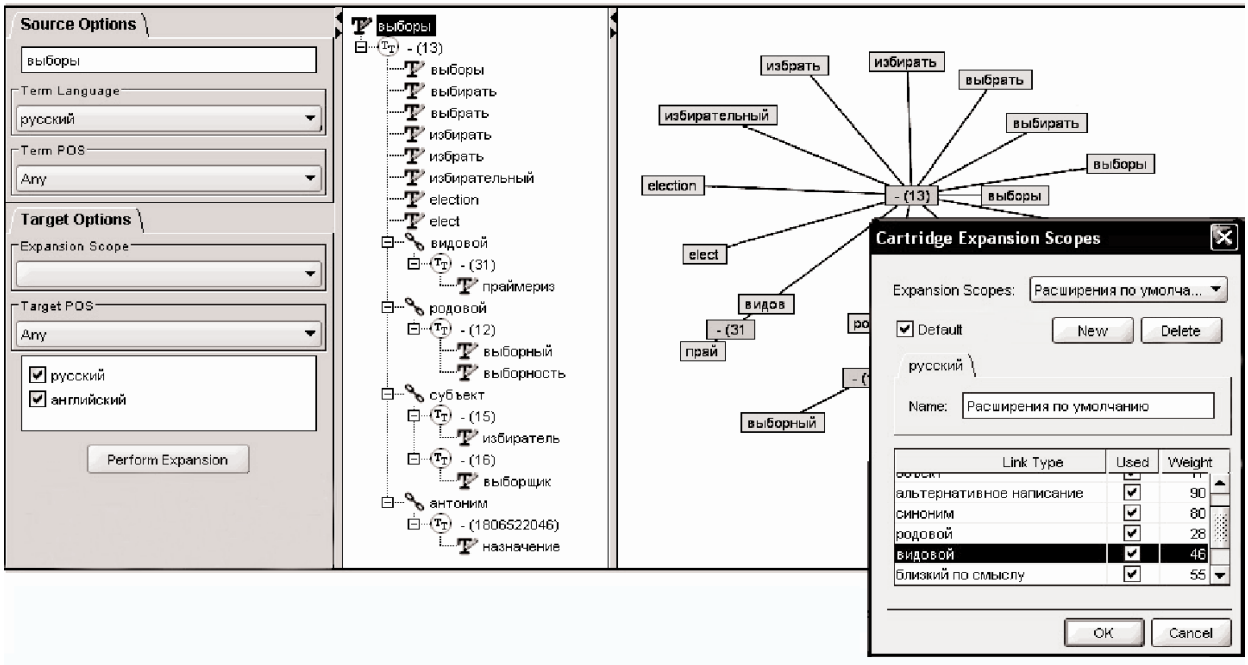

Рисунок 1. Фрагмент словаря RetrievalWare в интерфейсе Knowledge Workbench. Семантические отношения и их весовые коэффициенты.

В процессе интеллектуального анализа запроса для каждой лексической единицы запроса определяются:
- связанные понятия в доступных словарях;
- связанные с понятиями узлы таксономических иерархий;
- набор лексических единиц, находящихся на пересечении выявленных понятий.

Причинами, по которым пересечение может быть не найдено являются некорректный состав, или неполнота словарей, или неоднозначность условий запроса. В этом случае необходимо получить от пользователя дополнительную информацию через специальный интерфейс, в котором отображается неоднозначность, ее неприятные последствия и возможность их устранения путем выбора тех понятий, которые соответствуют информационной потребности, при этом:
- список понятий может быть ограничен только существующими в области поиска;
- в запросе могут быть выделены лексические единицы, связанные со статистически доминирующими понятиями, для замены их более точными;
- в случае отсутствия лексических единиц запроса в индексах могут быть предложены близкие им по написанию;
- интерфейс может содержать статистическую информацию о количестве соответствующих запросу документов и их тематическом спектре.

3. Применяемые лингвистические и статистические подходы

Качественный поиск возможен только при совмещении лингвистического и статистического подходов.

Создание качественных словарей является весьма трудоемким процессом, в котором доминирует ручной

труд и который, в нашем случае, состоит из следующих этапов:

1) на основе массива документов формируется набор наиболее значимых лексических единиц, собирается информация об их частотных характеристиках;

2) проводится морфологическая нормализация лексических единиц с определением частей речи;

3) экспертным путем лексические единицы связываются семантическими отношениями с понятиями, описывающими предметную область;

4) по возможности, на основе отношений типа род-вид и часть-целое понятия организуются в таксономические иерархии;

5) для каждого из семантических отношений вычисляется семантическая близость;

6) проводится верификация словаря как на внутреннюю непротиворечивость, так и на непротиворечивость с другими одновременно используемыми словарями.

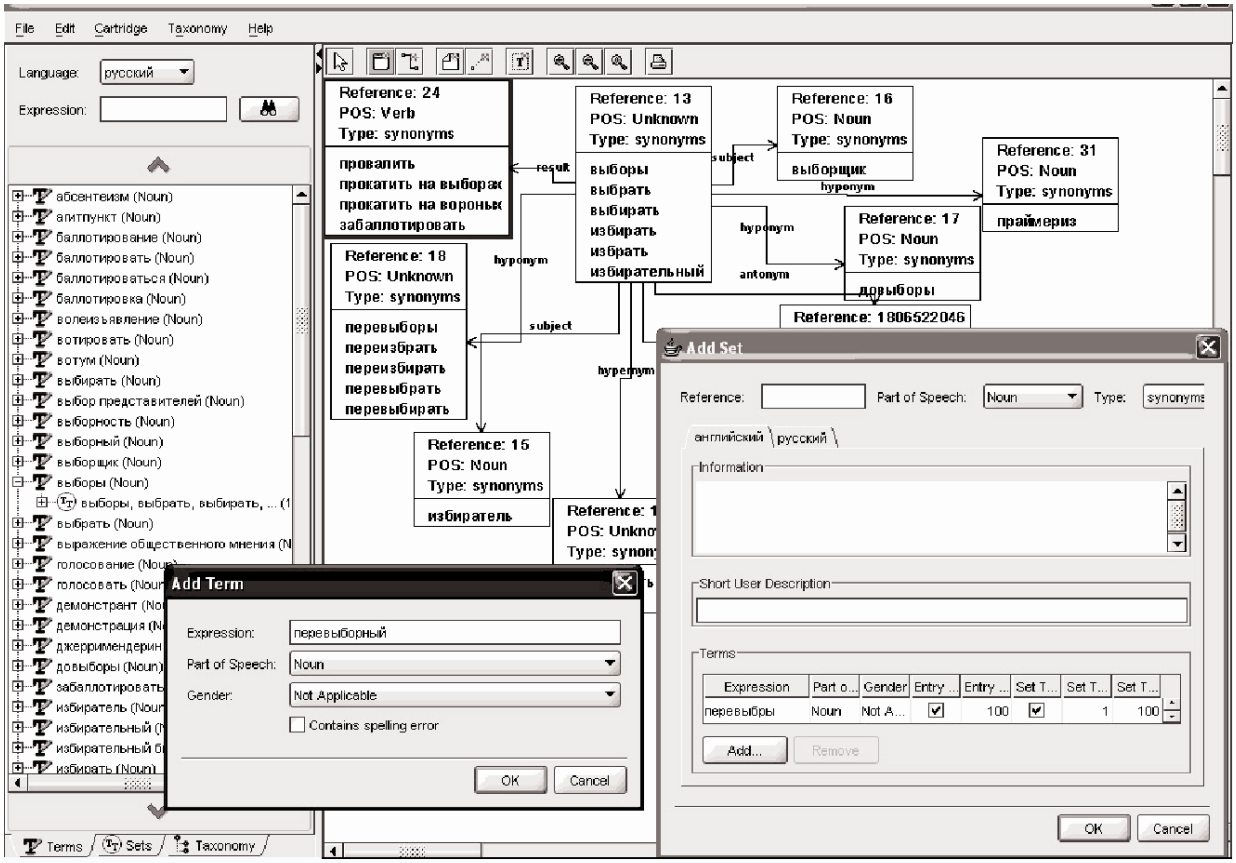

Рисунок 2. Создание словаря в Knowledge Workbenchh

На начальном этапе создания словаря, когда в распоряжении разработчиков есть только корпус текстов, хорошо зарекомендовали себя статистические методы, которые применяются компанией «Ретранс» для составления фразеологических словарей и описаны в книге [3]. Они позволяют создать фразеологический словарь предметной области, проверить его пересечение с соседними областями и построить иерархии уточнений многословных понятий.

После получения основы словаря проводится связывание лексических единиц и понятий, при этом желательно учитывать следующие аспекты:

1. При расчете семантической близости между понятием и лексической единицей следует учитывать не только их наличие в словарях синонимов, но и результаты анализа частоты соответствий между ними как в текстах по специальным тематикам, так и в текстах общей направленности.

2. Важна специфичность лексических единиц, которая непосредственно связана с числом связанных с ними понятий. В случае со словарем по предметной области все относительно просто, так как ее специфика служит естественным фильтром, позволяющим исключить нерелевантные понятия.

3. Учет сочетаемости слов в лексических единицах достигается специальной обработкой идиом, смысл которых разрушается при морфологическом изменении некоторых составных элементов. Наличие таких лексических единиц в словарях должно быть согласовано с применением алгоритма морфологического анализа.

Следует отметить, что идиомы в основном встречаются в общелексических словарях и практически отсутствуют (или могут быть морфологически нормализованы) в словарях по предметным областям. В последнем случае существенное увеличение точности поиска дает их организация в иерархическую структуру, начиная с парных словосочетаний с последовательным добавлением уточняющих слов.

4. Морфологическую нормализацию лексических единиц, в дополнение к сказанному, желательно реализовать в соответствии со следующим предпочтениями:

• отсутствие предлогов в многословных лексических единицах;
• учет порядка следования элементов многословных лексических единиц;
• учет различных вариантов написания, в том числе сокращений, инициалов и аббревиатур;
• учет использования прописных букв и других орфографических особенностей;
• учет транслитерации.

5. Следует учитывать однородность охватываемой словарем предметной области, которая может быть определена как общее распределение лексических единиц между понятиями. Эта характеристика рассматривается в совокупности с полнотой охвата предметной области. Необходимо учитывать, что различные сегменты предметной области могут быть описаны в словаре с разной глубиной и подробностью. Достижение приемлемого компромисса между этими характеристиками достигается за счет создания большого количества небольших словарей, охватывающих строго определенные предметные области, и предоставления пользователю возможности самостоятельно выбрать корректный их набор как в процессе индексации, так и при поиске.

Для тестирования словаря на исходном корпусе текстов и на текстах по смежным предметным областям мы используем инструментальное средство Knowledge Workbench.

4. Опыт разработки лингвистического процессора русского языка для системы RetrievalWare

Требования к полноте и точности морфологического анализа в поисковых системах зависят от обрабатываемого контента. В закрытых системах на уровне морфологической обработки текстов можно использовать модули детерминированного анализа с конечным объемом морфологического словаря, включающем и покрывающем всю лексику из индексируемых документов или, возможно, из семантических словарей, используемых для индексирования и классификации.

В открытых системах потенциально должны анализироваться любые лексические единицы, и хорошим решением проблемы может быть максимальное покрытие различных типов текстов (интернета, литературного языка и т.п.), охват как современной, так и исторической лексики, в том числе и случаев неправильного или нестандартного употребления (в интернет-чатах, блогах и т.п.). Поскольку в силу открытости языка это практически неосуществимо, приходится использовать дополнительные средства специализации и настройки, в том числе:

• расширенный анализ с распознаванием неправильно используемых, но распространенных форм слов (*ехай и т.д.);
• расширенный анализ с распознаванием слов с ошибками (*агенство и т.д.);
• анализ исторических форм слов;
• анализ вариантов написания слов (для фамилий, географических названий и т.п.).

В процессе разработки лингвистического процессора русского языка для RetrievalWare мы провели эксперимент по морфологической обработке массива практически всех доступных в электронном виде СМИ с 1994 по 2005 год. Из массива текстов объемом более 30 Гб было автоматически выделено для анализа около 8,6 млн. различных «словоформ», которые были отсортированы по частоте встречаемости. По нашим наблюдениям, около 30 % словоформ составили различные коды, части словоформ, иностранные слова и т.п. В эксперименте были задействованы три модуля морфологического анализа, все они использовали метод детерминированного анализа, при этом объемы морфологических словарей у каждого из них составляли не менее 200 тыс. слов. В результате оказалось, что каждый из модулей смог распознать от 15 до 17% представленных словоформ. Даже модуль с самым большим словарем не смог проанализировать большее количество фамилий и названий, которые обладают значительным семантическим весом в документах и в поисковых системах интенсивно используются в запросах. Другой недостаток практически всех испытываемых модулей – неполнота анализа, когда словоформа распознавалась и сводилась только к одному слову с определенным типом словоизменения, в то время как она могла быть сведена к двум и более; чаще всего это опять же касалось имен собственных (омонимичность имен - фамилий - названий городов - общих слов).

По нашему мнению максимальную полноту сможет обеспечить использование эвристического бессловарного анализа, т.е. возможности приведения всех или большей части словоформ к одной или нескольким «исходным» формам. В основе этого анализа могут лежать различные методики, чаще всего – метод аналогии в соче-

тании с предварительно созданной системой морфологической классификации слов. Данный метод не является самым оптимальным с точки зрения объемов подлежащих индексированию результатов анализа, количество которых можно сократить за счет дополнительного анализа контекста уже не только на уровне словоформы, но на уровне хотя бы синтагмы. Еще одним из способов оптимизации является сочетание детерминированного и эвристического методов анализа. Для разрешения омонимии имен собственных необходимо применение методов идентификации их классов с учетом контекста.

С 2004 года мы разрабатываем общелексический словарь русского языка. За эталон был взят аналогичный словарь английского языка, который давно применяется в RetrievalWare и создан на основе тезаурусов American Heritage, WordNet и Roget. В Таблице 1 приведена динамика наполнения нашего словаря.

Количественные характеристики	**Словарь английского языка**	**Словарь русского языка**		
		2004 г.	**2005 г.**	**2006 г.**
Лексических единиц (тыс.)	630	136	198	290
Количество понятий (тыс.)	100	31	41	58
Среднее кол-во «входов»	2	1,3	1,5	1,75
Число типов отношений	7	6	7	7
Среднее кол-во связей	6,2	4,4	4,8	5

Таблица 1

По результатам его эксплуатации мы сделали следующие выводы:

- Для общеупотребительной лексики неверно, что чем объемнее семантический словарь, тем лучшие результаты поиска он обеспечивает. Наращивание объема с определенного момента должно проводиться в основном за счет фразеологических словосочетаний, при этом соответствующим образом должен быть адаптирован алгоритм индексации и поисковый интерфейс. Важность включения в словарь многословных лексических единиц особенно важна в словарях по предметным областям (см. Рисунок 3)

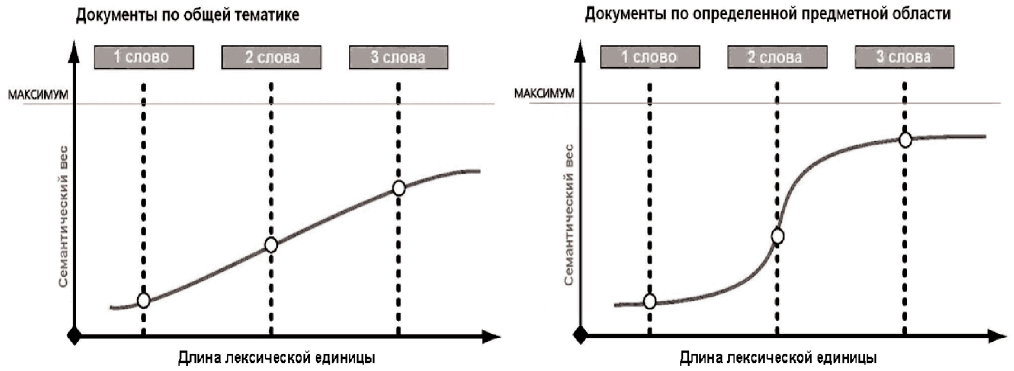

Рисунок 3. Зависимость семантического веса от длины лексической единицы

- При выборе слов-расширений запроса должны использоваться отношения, характеризующиеся оптимальной семантической близостью. При увеличении объема семантического словаря относительная семантическая близость входящих в его состав синонимов становится менее предсказуемой. Некоторые понятия могут оказаться связанными с очень близкими синонимами, в то время как другие – нет. Такая несогласованность приводит к проиллюстрированному на Рисунке 4 противоречию между полнотой и точностью, так как недостаточная семантическая близость увеличивает полноту при уменьшении точности, в то время как достаточная семантическая близость при увеличении полноты может даже привести к увеличению точности.

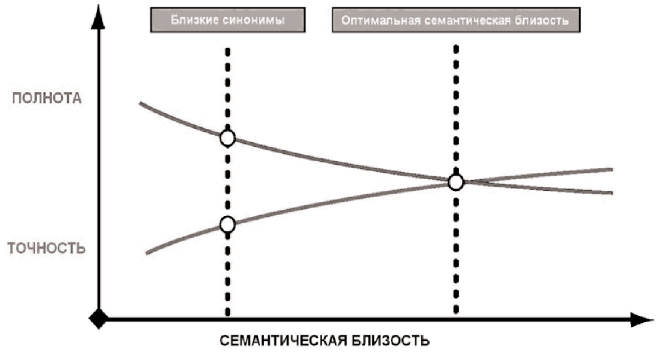

Рисунок 4. Влияние семантической близости на точность и полноту поиска

- Необходим приоритет применения разных типов словарей. В нашем случае наибольшим приоритетом обладают пользовательские словари, затем следуют словари по предметным областям, и только для обработки оставшихся лексических единиц применяется словарь общеупотребительной лексики.
- Необходимо применение алгоритмов идентификация специальных классов лексических единиц, таких как акронимы и аббревиатуры. Как правило, они точно соответствуют понятию, но по написанию часто совпадают с другими короткими словами, которые включаются в словарь стоп-слов.
- Среди различных понятий, соответствующих определенной лексической единице, часто присутствует наиболее употребимое. Эта особенность языка приводит к тому, что поиск документов с альтернативными понятиями становится очень затруднительным. Решение проблемы видится в использовании алгоритмов интеллектуально анализа запроса, для работы которых необходима информация о частоте встречаемости понятий и лексических единиц в массиве текстов. Структура семантических словарей должна быть приспособлена для хранения и использования такой информации.

Структура наших словарей позволяет создавать кросс-языковые семантические словари, и в настоящее время мы поставляем англо-китайский и англо-арабский словари, а также шестиязычный общелексический словарь для европейских языков. Из характеристик, которые приведены в Таблице 2 видно, что их объем во много раз меньше общелексических словарей исходных языков. Это объясняется тем, что кросс-языковые словари содержат только однозначные переводы, поэтому включение даже близких синонимов в словарные статьи практически невозможно. Сфера применения таких словарей – поиск в массивах документов на смешанных языках, а не семантическое расширение запроса. Наличие электронных словарей и программ-переводчиков с открытым интерфейсом позволяет эффективно автоматизировать создание кросс-языковых словарей.

Словарь	Языков	Понятий	Лексических единиц
Общеевропейский	6	172483	322905
англо-китайский	2	84968	199165
англо-арабский	2	67411	154685

Таблица 2.

5. Опыт разработки отраслевых словарей и таксономических классификаторов

Большинство отраслевых словарей содержат многословные лексические единицы и являются кросс-языковыми. В настоящее время поставляется более 40 словарей и более 70 классификаторов, охватывающих основные отрасли человеческой деятельности. В их основе лежат такие ресурсы как MESH, WAND, DTIC, Gene Ontology, Pro Quest, Access Innovations и другие. В Таблице 3 перечислены характеристики некоторых из них.

Словарь и источник	Языков	Понятий	Лекс. ед.	Корневых узлов	Уникальных лекс. ед.
AI News	1	5383	5422	6	5210
MeSH Chemicals and Drugs	1	8983	10196	27	6570
WAND Science and Technology	9	1409	18414	24	14354
PQ U.S. Government	1	1958	1313	77	928

Таблица 3.

В таблице 4 представлены характеристики некоторых отраслевых словарей, разработанных нами для русского языка.

Словарь и источник	Языков	Понятий	Лекс. единиц	Уровней иерархии	Слов в лекс. ед.
География мира	2	0,7 тыс.	2,8 тыс.	6	1,9
География России	1	1,2 тыс.	1,3 тыс.	4	1,2
Органы госвласти	1	3,8 тыс.	16,5 тыс.	8	6,6
Общество и политика	1	2,5 тыс.	6 тыс.	3	2,3
Российские компании	1	0,8 тыс.	1,1 тыс.	2	1,9

Таблица 4.

Выводы

Если обобщить наш опыт, то можно сделать следующие выводы:

1. Качественные и сравнительно недорогие словари формируются с привлечением статистического анализа корпуса текстов. Работу эксперта по добавлению в словарь тысячи понятий мы оцениваем в 1-1,5 человеко-месяца.

2. Наиболее трудоемким является создание таксономических иерархий. Для этих целей мы используем существующие классификаторы и справочники. Для ряда предметных областей приемлемые результаты получаются переводом иностранных таксономий. К сожалению, ряд наиболее востребованных предметных областей, таких как юриспруденция, оборона и безопасность, финансы и кредит требуют привлечения соответствующих специалистов.

3. Мы прекратили дальнейшее пополнение общелексического словаря русского языка, и в настоящее время сосредоточили усилия на его уточнении, в основном путем исключения полисемии с распространенными именам собственными.

4. Алгоритм морфологического анализа должен учитывать лексический состав семантических словарей. Для обработки имен собственных необходимо привлечение эвристических алгоритмов с правилами разрешения полисемии.

5. Довольно много ресурсов расходуется на разработку словарей, которые уже существуют на рынке. Продуктивной работе могло бы поспособствовать создание единого реестра лингвистических ресурсов, разработанных отечественными научными учреждениями и коммерческими организациями.

Список литературы

1. Чубинидзе К.А. Использование технологии динамической классификации для интенсификации аналитической деятельности // М.: ИТТП. 2005. № 3, С.56-65.
2. Громов А.И., Чубинидзе К.А. Управление знаниями и семантический анализ текстов в системе RetrievalWare компании Convera // М.: ИТТП. 2005. № 3, С.37-56.
3. Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А., Компьютерная лингвистика и перспективные информационные технологии // М.: Русский мир. 2004. С.