Стокона на РОМИП-2006

© А. Огарок

Научно-производственная фирма Стокона
Ogarok_AL@stocona.ru

Аннотация

Статья посвящена описанию итогов участия научно-производственной фирмы Стокона в дорожке вопросно-ответного поиска на семинаре РОМИП-2006. Приведено краткое описание методов лингвистического анализа и оценки релевантности, реализованных в системе Stocona Search.

1. Введение

Целями участия научно-производственной фирмы Стокона в семинаре РОМИП-2006 являлись: расширение круга вопросов, рассматриваемых на семинарах РОМИП за счет включения в него вопросно-ответной дорожки, а также исследование разработанных в компании технологий поиска информации на коллекции русскоязычных текстов. В данной работе дано краткое описание использующихся для этой цели методов лингвистического анализа, реализованных в системе Stocona Search. Проведена оценка эффективности реализованных методов поиска на основе анализа результатов выполнения заданий вопросно-ответной дорожки РОМИП-2006, а также сравнения с результатами тестирования системы Stocona Search на заданиях конференции TREC.

2. Вопросно-ответный поиск

2.1 Постановка задачи

Вопросно-ответный поиск был введен в состав дорожек семинара РОМИП-2006 по предложению научно-производственной фирмы Стокона. Организаторы предложили участникам список из 615 вопросов, на которые необходимо было найти ответы в

исходном наборе данных коллекции Narod.ru. Тестовые случаи относились к следующим типам: вопросы к определению, к подлежащему, к прямому дополнению, к обстоятельству, к косвенному дополнению, к прямому дополнению [1].

Оценки асессоров использовались для вычисления значений показателя Mean Reciprocal Rank (MRR) по двум вариантам метрик семинара РОМИП-2006:

- метрика romipReciprocalRank (семинара РОМИП) для которой коэффициенты десяти мест наиболее релевантных ответов определяются линейкой - {1; 0,9; 0,8; 0,7; 0,6; 0,5; 0,4; 0,3; 0,2; 0,1});
- метрика trecReciprocalRank (конференции TREC) для которой коэффициенты пяти мест наиболее релевантных ответов определяются линейкой - {1; 0,5; 0,33; 0,2; 0,1}).

Также оценивались показатели общего числа полученных ответов, количества точных и приблизительных ответов, а также точность ответов.

2.2 Описание прогона

Тестовая коллекция из 615 вопросов была разделена организаторами семинара РОМИП на 4 равные части. Каждая последующая часть вопросов выдавалась участникам вопросно-ответной дорожки после получения результатов тестирования по предыдущему набору вопросов.

Результаты тестирования системы Stocona Search были получены в ходе одного прогона по каждому представленному набору вопросов. Данные результаты были переданы для оценок асессорами.

Словарные базы системы Stocona Search не адаптировались под тексты коллекции Narod.ru, а применялись в стандартном варианте, используемом при тестировании Stocona Search на вопросах конференций TREC 2001 - 2006.

Для тестирования использовалась стандартная конфигурация системы Stocona Search, реализующая индексацию текстов с учетом морфологии и синтаксиса, а также учет семантики слов на этапе оценки лингвистической релевантности.

2.3 Описание метода

Методической основой организации процесса поиска в системе Stocona Search является полный цикл лингвистического анализа индексируемых текстов и запроса пользователя, а также

ранжирование результатов поиска по степени синтактико-семантического соответствия найденных предложений проиндексированных текстов предложению запроса.

Исследуемый метод вопросно-ответного поиска состоял следующих этапов:

1. Формирование поискового шаблона и составление списка группировок множества вхождений слов запроса в тексты документов.

2. Оценка синтактико-семантического соответствия вопроса пользователя предложениям текстов из списка, составленного на первом этапе.

На первом этапе индексный процессор отбирал участки текстов, содержащие ключевые слова тестового вопроса с учетом лингвистических и статистических критериев. Отбираемые участки текста упорядочивались по рассчитанному коэффициенту базовой релевантности. Количество отбираемых участков текстов ограничивалось 100 текстами по 5 наиболее релевантных абзацев в каждом из них.

На втором этапе лингвистический процессор проводил синтактико-семантический разбор отобранных участков текстов и ранжирование результатов поиска по коэффициенту лингвистической релевантности.

При поиске учитывались морфология, синтаксис, семантика слов, синонимы и родственные слова запроса пользователя и текстов, проводилась проверка наличия смыслового ответа на вопрос пользователя в смежных предложения (разрешалась анафорическая связь). Лингвистический анализ Stocona Search использует набор характеристик лексем, которые позволяют описать морфологическую, синтаксическую и семантическую структуру предложения.

Лексический и частично семантический анализ в системе Stocona Search обеспечивает выделение токенов и семантических классов. Он построен на использовании регулярных выражений и конечного автомата поиска паттернов в последовательности символов. Система управления базами данных построена на алгоритмах хеширования. Для задач сортировки и поиска используются алгоритмы обработки B-деревьев. Основой синтаксического и семантического анализатора является метод информационной доски. Основу метода составляют три элемента: информационная доска, совокупность источников знаний и управляющий этими источниками контроллер [2, 3]. В качестве источников знаний в

модулях синтаксического и семантического анализа выступают синтаксические и семантические правила продукции. Метод информационной доски дает хорошие показатели с точки зрения скорости и удобства построения парсеров естественного языка.

Теоретической основой для построения синтаксического анализа явилась грамматика зависимостей, где отсутствуют нетерминальные символы и главным членом предложения является глагол или глагольное ядро предложения, причем правила являются контекстно-зависимыми.

Совокупность правил продукции составляет формальную грамматику языка, однако грамматики могут быть использованы и для описания частных языковых явлений. Так, в системе используется грамматика для порождения римских и арабских числительных, которая обеспечивается сопоставление римских, арабских чисел, и чисел в литеральном представлении.

Методическим приемом, позволившим сделать теоретическое описание синтаксического и семантического уровня независимым от машинного анализа, явилось хранение синтаксических и семантических правил продукции в базе и использование универсального парсера, работающего с формальным представлением текста. Это позволяет структурировать систему, значительно сократить время, требуемое на перенастройку системы на новый язык и передать составление описания языка в руки экспертов-лингвистов. Семантические классы словарной подсистемы Stocona Search являются объектами онтологий.

Ранжирование ответов в выдаче осуществлялось на основе вычисленных значений базовой и лингвистической релевантности. Определяющее значение имело значение лингвистической релевантности. При ее расчете учитывалось: группировка слов в предложении, связь с заголовками текста, синтаксические и семантические роли слов.

2.4 Результаты оценки

Результаты оценки асессорами ответов участвующих в вопросно-ответной дорожке поисковых систем представлены в таблице.

Оценка результата Trec Reciprocal Rank показывает, что система Stocona Search при тестировании по вопросно-ответной дорожке семинара РОМИП показывает результаты, аналогичные полученным при ее тестировании на вопросах конференций TREC 2001 - 2006. Тестирование на коллекции TREC [4] проводилось ранее на стендах научно-производственной фирмы Стокона.

Анализ ответов системы по различным категориям запросов показал, что система Stocona Search хорошо отвечает на вопросы, направленные на поиск слов ответа, принадлежащих к учитываемым семантическим категориям (время, место, лицо, организации и т.п.).

Таблица

Показатель	**Значение**
Общее количество тестовых случаев вопросно-ответного поиска	615
Количество случаев, в которых нет ни одного ответа	146
Количество случаев, в которых нет точных релевантных ответов, но имеются частично релевантные ответы	226
Всего ответов	5509
Количество точных релевантных ответов	880
Количество частично релевантных ответов	158
Точность поиска (micro)	0,19
Точность поиска (macro)	0,17
Trec Reciprocal Rank	0.41
ROMIP Reciprocal Rank	0.48

В состав тестовых случаев были включены вопросы, ответы на которые должны были найдены с учетом синонимов и родственных слов. Практическая проверка использования данного механизма позволила убедиться в его эффективности. Использование классов синонимов и родственных слов существенно расширяет полноту поиска. Для обеспечения точности поиска необходимо применять набор тематически структурированных синонимических словарей.

Экспериментальная оценка показала, что исследуемые методы вопросно-ответного поиска являются достаточно универсальными и могут использоваться при создании широкого класса систем обработки текстовой информации.

Высокая эффективность реализованных в системе Stocona Search лингвистических алгоритмов обусловливает целесообразность дальнейшей ориентации алгоритмов работы поисковых систем на углубление роли синтактико-семантического анализа при разборе запроса пользователя и оценке лингвистической релевантности предварительно отобранных предложений ответов.

Считаем целесообразным при проведении вопросно-ответной дорожки семинара РОМИП-2007 ужесточить критерии оценки систем для более наглядного представления различий при использовании традиционного поиска по ключевым словам и лингвистического подхода. Предлагаем для этого в критерий оценки

«правильности» ответа ввести указание поисковой системой слова семантического ответа на вопрос пользователя.

Дальнейшими направлениями наших исследований являются: совершенствование алгоритмов синтактико-семантического анализа, использование расширенного состава онтологий и баз знаний для верификации ответа поисковой системы на запрос пользователя.

4. Заключение

Семинар РОМИП динамично развивается. В 2006 году он пополнился вопросно-ответной дорожкой, аналогом которой можно считать QA тесты конференции TREC на английском языке. Научно-производственная фирма Стокона в силу организационных причин не смогла принять участие в ряде других дорожек. Было бы весьма интересно принять нам участие в дорожке фактографического поиска. Надеемся принять участие в большем количестве дорожек конференции РОМИП-2007.

Результаты тестирования систем весьма полезны для оценки достигнутого уровня разработок и определения задач на будущее.

В заключение хотелось бы поблагодарить Игоря Некрестьянова и его коллег за самоотверженную работу по организации семинара и проведению оценок, а также Илью Сегаловича и компанию Яндекс за предоставленные коллекции документов.

Литература

[1] ROMIP Web site, 2006. http://romip.narod.ru
[2] Гради Буч. Объектно-ориентированный анализ и проектирование. Изд.: БИНОМ, 1999, с. 560.
[3] Englemore, R. and Morgan, T. 1988. Blackboard Systems. Wokingham, England: Addison-Wesley, p.v.
[4] TREC Web site, 2006. http://trec.nist.gov/

Stocona at ROMIP 2005

© A. Ogarok
Ogarok_AL@stocona.ru

The article presents description of Stocona's participation result in question-answering search in ROMIP seminar 2006 and contains brief description of linguistic analysis methods realized in the Stocona Search system.