ОБРАБОТКА ТЕКСТА С ПОМОЩЬЮ ТОМИТА-ПАРСЕРА

Р.В. Огородник, Л.В. Серебряная
Кафедра программного обеспечения информационных технолгий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: aharodnik@gmail.com, l_silver@mail.ru

Сложившаяся в современном информационном пространстве ситуация требует увеличения скорости и качества поиска информации из того многообразия, которое сейчас представляет информационное общество. Для решения задач интеллектуальной обработки информации и данных необходимо переработать структуру данных текста таким образом, чтобы извлечение знаний из текста происходило наиболее быстро и точно.

Введение

Одной из задач, которую необходимо решить при конструировании модели обработки текстовой информации и выделения из неё знаний это разбор предложений и построение структур данных, необходимых для следующего этапа разработки. Для этого был применён алгоритм Томита-парсера, с улучшениями, которые и рассмотрены в данном докладе.

I. Томита-парсер

В основу Томита-парсера лег GLR-алгоритм — это расширенная версия алгоритма LR-парсинга. Алгоритм выдает результаты своей работы в режиме реального времени, по мере продвижения вглубь текста, другие алгоритмы обработки естественного языка такой особенностью не обладают. Однако одного лишь алгоритма для полноценного анализа текста и извлечения из него структурированной информации недостаточно. Нужно учитывать морфологию и синтаксис языка обрабатываемого текста, подключить необходимые словари, понятные парсеру. Томита-парсер разрабатывался специально с прицелом на упрощение работы с алгоритмом. Был составлен несложный синтаксис для создания словарей и грамматик, продумана работа с морфологией русского языка. В минимальной конфигурации парсеру на входе отдается сам анализируемый текст, а также словарь и грамматика. Объем словаря и сложность грамматики зависят от целей анализа: они могут быть как совсем маленькими, так и огромными. Файл грамматики состоит из шаблонов, написанных на внутреннем языке/формализме Томита-парсера. Эти шаблоны описывают в обобщенном виде цепочки слов, которые могут встретиться в тексте. Кроме того, грамматики определяют, как именно нужно представлять извлеченные факты в итоговом выводе. В словарях содержатся ключевые слова, которые используются в процессе анализа грамматиками. Каждая статья этого словаря задает множество слов и словосочетаний, объединенных общим свойством. Например, «все города Беларуси». Затем в грамматике можно использовать свойство «является городом Беларуси». Слова или словосочетания можно задавать явно списком, а можно «функционально», указав грамматику, которая описывает нужные цепочки. Процесс обработки информации будет состоять из следующих этапов:

- Морфологический анализ слов;
- Синтаксический анализ текста;
- Построение графа связанности текста;
- Редактирование дерева слов;
- Наполнение семантической сети.

Томита-парсер призван решать задачи первого и второго этапов данного алгоритма.

II. Этапы работы

С помощью словаря форм определяется морфологическая роль слова в предложении. Основной проблемой, с которой пришлось столкнуться на данном этапе, являлось определение части речи и, соответственно, других морфологических значений. В случае омонимии определялась форма слова по частоте встречаемости инфинитива в данном тексте, в первую очередь и по тематике текста, если однокоренных искомому слову-омониму в тексте не найдено. Так как первый этап является подготовительным для дальнейших, то основной упор в нём делается на правильное определение частей речи и, на основе словаря инфинитива этого слова. Все служебные части речи из текста на этом этапе изымаются. Частицы и союзы участвуют и в дальнейших этапах метода, но для них не определяется форма, только определяется часть речи, чтобы иметь возможность наиболее точно разбить сложные предложения на логические и синтаксические части. Служебные части речи только помечаются, фильтрация их будет производиться впоследствии. На втором этапе производится синтаксический анализ текстов на основе уже выполненного морфологического разбора, где определяется роль слова в предложении и связи между словами. Определенным морфологическим и грамматическим ролям слов соответствует синтаксические роли, которые с более высокой точностью определяются, в случае если предлоги объединяются с существительными, к которым они отно-

сятся. Также на втором этапе проходит склеивание сущностей, в частности это делается с помощью словаря синонимов и с помощью переопределения местоимений в тексте на другие части речи, которые уже применялись в предыдущих предложениях. Здесь происходит разбиение сложных синтаксических конструкций: сложносочинённых и сложноподчинённых предложений на простые для лучшей реализации дальнейших построений на основе предложений на следующем этапе. Простейшие конструкции были разложены с помощью Томита-парсера, и были произведены доработки по улучшению алгоритма.

III. Доработанные элементы Томита-парсера

К недостаткам полученной семантической сети можно отнести несовершенство её лингвистического аспекта: все представляемые данные на естественном языке должны соответствовать заранее предопределенному шаблону. В ходе работы сети также рекомендовался шаблон вопросов, для недопущения введения в систему ложных знаний из-за возможных ошибок при разборе текстовой информации. Также, к недостаткам семантической сети следует отнести то, что в ходе работы не предлагалось исправлять орфографические и грамматические ошибки. Поэтому дополнительным требованием к эксперту было правильное с точки зрения грамматической и орфографической точки зрения построение предложений.

Лингвистический аспект получения знаний был разделен на две проблемные категории: морфологический и синтаксический. Морфологический аспект включает в себя распознование слов, словоформ, определение на основе грамматического разбора слова его формы и возможной. Опеределение возможной части речи для последующего связывания слов при синтаксическом разборе возможно уже на этапе морфологического анализа. Синтаксический анализ предложения с определенением роли слова в предложении помогает определить связи и понять структуру предложения и отношения слов между собой, для последующего построения соотношений. В качестве хранилища слов и словоформ предполагается использовать префиксные деревья для словооснов и смешанный вид с привлечением списков и деревьев для хранения возможных грамматических форм слова – префиксов и окончаний. Преимущество такого подхода в том, что не хранятся огромные цепочки повторяющихся символов для одного и того же слова, стоящего в разных формах. Таким образом, уменьшается размер файла со словарем. К примеру, для русского языка это дерево вместе с дополнительными данными занимает порядка 2 мегабайт. Скорость поиска по такому дереву значительно выше, чем по отсортированному файлу. Теперь она зависит не от размера всего словаря, а только от того, какой длины слово ищется.

В качестве развития данной семантической сети предполагается не только устранение уже существующих недостатков, но и внедрение в систему скрытых структур знаний на основе обобщений уже полученных данных. Имплиционные знания будут определяться экспертом на основе обобщений имеющихся структур, позже, во время работы, система сама будет определять эти структуры и делать на основе их аналитические, дедуктивные, индуктивные и другие методы получения знаний.

IV. Выводы

Данное исследование ставит своей задачей разработку базовых структурированных данных для наполнения ими семантической сети. Наполнение будет происходить с экспертом, для достижения большей точности извлекаемых знаний. Связи сети можно будет разделить на категории: определяющие тип объектов, количественные, временные, атрибутивные и логические. Функциональные связи ввиду их большого количества будут вынесены в структуру данных, что ограничивает возможности поиска. Однако наложение дополнительных связей, определяющих тип объекта для этих данных, позволяет пользоваться ими при поиске правильного ответа с высокой степенью свободы. Пространственные связи используются только в качестве возможности их объективного определения (например, «граничит» вместо «недалеко»), чтобы избежать субъективной оценки, как на этапе обучения системы, так и в ходе работы с ней. Таким образом, предложенную семантическую сеть можно классифицировать как неоднородную, бинарную сеть с вышеперечисленными типами отношений. В качестве развития данной семантической сети предполагается внедрение в систему скрытых структур знаний на основе обобщений ранее полученных данных. Имплиционные знания будут определяться экспертом на основе обобщений имеющихся структур. Затем во время работы система сама определит эти структуры и построит на их основе различные аналитические методы получения знаний.

1. Yandex. [Электронный ресурс] / Что такое Томита-парсер, как Яндекс с его помощью понимает естественный язык, и как вы с его помощью сможете извлекать факты из текстов . – Режим доступа: http://habrahabr.ru/company/yandex/blog/219311/ – Дата доступа: 01.09.2014.
2. Тельнов Ю. В. Интеллектуальные информационные системы. М., 2004. С.67–69
3. Yandex - [Электронный ресурс] /Томита-парсер. Учебник. – Режим доступа: http://api.yandex.ru/tomita/doc/tutorial/concept/about.xml/ – Дата доступа: 01.09.2014.