Речь на защите выпускной квалификационной работы бакалавра

Уважаемая аттестационная комиссия! Вашему вниманию представлена выпускная квалификационная работа на степень бакалавра по специальности «Информационные системы», выполненная студентом группы Фт-47081 Усталовым Дмитрием. Тема работы: «Исследование и разработка системы автоматического извлечения ключевых фраз из текста на естественном языке».

Итак, что же такое «ключевая фраза»? Ключевая фраза — это выражение, состоящее из одного или нескольких ключевых слов, представляющее собой важнейший информационный сегмент документа.

Сегодня интеллектуальные информационные системы нашли широкое применение в области здравоохранения. Одной из важнейших составляющих современных медицинских информационных систем является подсистема интеллектуального анализа текстовых данных. Адекватность её функционирования напрямую зависит от качества работы модуля извлечения ключевых фраз.

В общем виде, задача возникает в библиотечном деле, лексикографии и терминоведении, а также в информационном поиске. Ключевые фразы могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, классификации и визуализации.

К сожалению, ведущие системы разработаны на Западе и ориентированы исключительно на обработку западноевропейских языков , что делает их непригодными для обработки текстов на русском языке. Системы, разработанные в России также непригодны для использования из-за устаревших и неэффективных методов извлечения ключевых фраз, а также ограничительных условий распространения.

Аналоги

Существует большое количество систем автоматического извлечения ключевых фраз из текста на естественном языке . Их список, надеюсь, прекрасно виден на экране. Так вот, необходимая система оценивается по следующим практически важным критериям :

1. поддержка русского языка;
2. качество результата работы по итогам экспертной оценки ;
3. доступность;
4. независимость аналога от наличия онтологии заданной области знаний или специализированного тезауруса.

В результате сравнения аналогов по обозначенным критериям целесообразно выбрать аналог №7 (**АОТ**) в качестве прототипа.

Прототип

Сейчас на экране представлена структурная модель прототипа. Как видно, в систему поступает исходный текст на русском языке, который проходит три этапа обработки. Сначала текст разбивается на значимые блоки, после этого выполняется морфологический анализ входящих в него слов, а в конце концов на основании выполненных этапов графематического и морфологического разбора строятся синтаксические отношения между словами. Результатом работы прототипа является дерево синтаксического разбора каждого предложения, входящего в исходный текст.

На самом деле, функционирование прототипа можно показать более наглядно. Поскольку в названии нашей специальности где-то фигурировало слово «медицина», я зашёл в «Википедию» и в течение пятнадцати минут нажимал на кнопку «Случайная статья» в ожидании любой около-медицинской статьи. Не знаю, быть может, это какая-то форма гадания, я такими вещами не занимаюсь, но «Википедия» мне выдала статью про синдром Дауна.

Давайте посмотрим что делает прототип с фрагментом текста этой чудесной статьи. На этапе графематического анализа, как я уже говорил, текст разбивается на абзацы, предложения, отдельные слова и символы. После этого проводится морфологический анализ каждого слова. Например, слово «синдром» в нашем случае определяется как имя существительное мужского рода в единственном числе и именительном падеже. После построения морфологической интерпретации, система ищет согласования слов друг с другом, формируя синтаксические группы. Группы, в свою очередь, связываются друг с другом, образуя дерево.

Именные группы, выделенные при помощи синтаксического анализатора, можно считать ключевыми фразами исходного текста.

Поговорим о критике прототипа. У нас есть два пункта:

1. применённый в прототипе метод распознавания ключевых фраз обладает недостаточной точностью : после обработки текста мы имеем достаточно большое количество именных групп и мы не можем судить о том, какая из выделенных групп является наиболее важной в данном тексте;
2. стоит отметить, что качество работы современных морфологических анализаторов превосходит морфологический анализатор АОТ.

Предлагаемое решение

Само собой, критику надо преодолевать. Для этого были приняты меры:

1. вычислять статистическое значение терминологичности каждой извлечённой именной группы и на основе этого значения ранжировать полученный список именных групп;
2. заменить морфологический анализатор АОТ на одно из современных решений.

Предлагается вычислять значение терминологичности при помощи метрики C-value. Расчётная формула приведена на экране. Легко видеть, что чем больше частота встречаемости термина—кандидата в тексте и чем выше его длина, тем больше его вес в исходном тексте. Однако если этот кандидат входит в большое количество других словосочетаний, то его вес уменьшается. Путём сортировки списка кандидатов в термины по убыванию значения C-value можно получить список ключевых фраз, наиболее адекватных исходному тексту. Кстати, аналогичное решение применено в системе TerMine, получившей наивысшую оценку качества результата работы среди систем, не использующих онтологию заданной области знаний в процессе выделения ключевых фраз.

Назовём блок, вычисляющий C-value, блоком выделения ключевых фраз. На экране показана его структурная модель. На входе этого блока имеется дерево синтаксического разбора исходного текста. Мы явным образом фильтруем из исходного дерева список именных групп, вычисляем для них C-value по вышеприведённой формуле, приводим именные группы в каноническую форму, и сортируем полученный список ключевых фраз по убыванию C-value.

Применение C-value позволит более значимым ключевым фразам оказаться в начале списка, а менее значимым — в конце. Теперь разберёмся с морфологией. Определим набор критериев оценки морфологических анализаторов:

1. поддержка русского языка;
2. возможность определения части речи и грамматических характеристик слова;
3. возможность выделения основы слова;
4. качество анализа по итогам экспертной оценки ;
5. доступность.

В результате сравнения морфологического анализатора АОТ с

аналогами, целесообразно выбрать аналог №4 (**myaso**) в качестве готового морфологического анализатора.

Структурная модель предлагаемого решения представлена на экране. Как видно, в структуру прототипа внесён блок выделения ключевых фраз.

Проектирование решения

В результате проектирования предлагаемого решения разработано Техническое задание на 16-ти листах. Для разработки выбрана платформа Rubinius — перспективная реализация языка программирования Ruby на основе виртуальной машины LLVM.

Система построена на основе архитектуры REST и стандартных протоколов JSON и XML. Также предусмотрена возможность запуска системы в «облачной» среде. Конечно же, система называется „Tesuçk”. Почему бы и нет?

Главная страница Web-интерфейса выглядит примерно так, как показано на экране. В большое поле ввода вставляется исходный текст, выбирается формат выдачи результата и нажимается кнопка «Обработать текст». Здесь очень плохо видно, но извлечённые системой ключевые фразы действительно отсортированы по убыванию C-value и оформлены в виде таблицы. Кроме того, система неплохо документирована и её программный интерфейс доступен любому разработчику, способному работать с форматами JSON или XML.

Мне приятно сообщить, что система Tesuçk полностью построена на основе свободного программного обеспечения. В процессе выполнения инженерной реализации мне пришлось дописывать пару существующих Open Source-решений и вносить в них необходимую мне функциональность. Внесённые изменения поддержаны авторами оригинальных продуктов и переданный код востребован сообществом.

Внедрение

Очень важно не только реализовать предлагаемое решение, но и сделать его нужным, а также представить его людям. На сегодня я добился следующих успехов:

1. система Tesuçk развёрнута в PaaS-облаке Cloud Foundry, территориально расположенном в ЦОД ИММ УрО РАН и доступна по URL, указанному на экране;
2. ведутся работы по интеграции Tesuçk в популярный сервис коллективных переводов translated.by по договорённости с компанией JetStyle.

Дальнейшая работа

В будущем, я планирую заниматься развитием этой системы в рамках создания облачного сервиса автоматической обработки текста. В частности, мне хочется отметить три важных момента:

1. на днях я выиграл грант на оплату оргвзноса и проживание в Петербурге во время прохождения российской летней школы по информационному поиску в августе этого года;
2. я прекрасно понимаю роль различных свидетельств об интеллектуальной собственности в современном мире и в ближайшее время займусь решением вопроса о надлежащем оформлении моих разработок;
3. мне понравилось заниматься наукой, я очень благодарен своему руководителю, консультантам, рецензенту и многим другим людям, которые помогали мне и поддерживали меня в процессе работы над системой Tesuçk. В магистратуре я продолжу заниматься компьютерной лингвистикой и хочу на основе своих разработок создать систему синтеза изображения по тексту.

На этом моя презентация заканчивается. Если есть минута времени, то я могу показать видеозапись «живой» работы с системой.

Ответ на вопрос рецензента

Почему в медицинских информационных системах адекватность функционирования подсистемы интеллектуального анализа текстовых данных зависит от качества работы модуля извлечения ключевых фраз?

Основной задачей интеллектуального анализа текстовых данных является обнаружение знаний в неструктурированном тексте на естественном языке. Как правило, в процессе анализа выполняется построение специализированного тезауруса или онтологии области знаний, применение которых определяется конкретной прикладной задачей, например — семантическим поиском информации.

Вершинами онтологии являются концепты, то есть термины заданной предметной области. Таким образом, чем точнее мы можем получить список терминов исходного текста, тем корректнее мы можем построить онтологию: её концепты и отношения между ними. Каждое «ложное» срабатывание модуля выделения ключевых фраз создаёт фрагменты онтологии, неадекватные исходному тексту, что снижает эффективность любых запросов к системе знаний.

С ростом популярности и сложности систем, основанных на знаниях, проблема точности выделения ключевых фраз принимает всё более серьёзное значение в том числе и в медицинских системах, например в информационно-справочных или же консультативно-диагностических системах.