ОСОБЕННОСТИ СЕТЕВОГО АНГЛОЯЗЫЧНОГО ЛИНГВИСТИЧЕСКОГО ПРОЦЕССОРА ДЛЯ ФОРМАЛИЗАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

FEATURES OF THE NETWORK ENGLISH LINGUISTIC PROCESSOR FOR FORMALIZATION OF THE TEXT INFORMATION IN A NATURAL LANGUAGE

А.А. Петров *(info@apetrov.ru),*
МТУСИ

И.П. Кузнецов *(igor-kuz@mtu-net.ru),*
ИПИ РАН

В докладе рассматривается лингвистический процессор для формализации англоязычной текстовой информации на естественном языке как сетевой компонент Интернет-проекта. Рассматриваются задачи лингвистического процессора, особенности его англоязычной версии, а также сетевая интеграция в Интернет-портал.

Введение

Проблема автоматической обработки автобиографических сведений (резюме), написанных в свободной форме, является важной задачей для кадровых и рекрутинговых агентств. Такие сведения имеют вид текстов на естественном языке и содержат типовой набор данных о человеке. У большинства соискателей такие тексты уже имеются в электронном виде.

Но использовать их, например, для поиска нужного работника, оказывается сложной процедурой. Поэтому требуется их предварительная формализация для автоматического ввода в существующую информационно-поисковую систему. В качестве такой системы был выбран крупнейший Интернет-портал HeadHunter.ru.

Данный Интернет-ресурс работает не только с российскими, но и зарубежными компаниями, предоставляя им информацию на английском языке.

Для соискателей, желающих получить работу, Headhunter.ru решил повысить уровень удобства работы с системой и разработать новый сервис: кандидат пересылает имеющееся у него резюме в виде текста через форму на сайте.

Существующие программные решения на тот момент, такие как KeyStaff Solutions, были основаны на частотном анализе текстов и не давали требуемых результатов для английского языка.

Поэтому для обработки данной текстовой информации (автоматического преобразования в формат сайта) было решено разработать систему LINGVO-MASTER.

1. Задачи лингвистического процессора

Система LINGVO-MASTER содержит семантико-ориентированный лингвистический процессор (ЛП) [1], который управляется с помощью лингвистических знаний (ЛЗ). В его основе использованы методики, основанные на технологии баз знаний (БЗ).

Каждое резюме содержит типовой набор данных о человеке (ФИО, год рождения, домашний адрес, время и место учебы и др.). Эти информационные объекты зачастую грамматически не согласованны между собой. Требуется, во-первых, выделить эти объекты из текста заявки, во-вторых, привести к единообразному виду, и, в-третьих, связать их между собой, например, время с местом учебы или работы.

Выделение информационных объектов осуществляется ЛП, который состоит из оболочки, управляемой лингвистическими знаниями (ЛЗ). Настройка на выделяемые объекты и анализируемые формы языка сводится к разработке соответствующих ЛЗ.

2. Особенности англоязычной версии процессора

Многие информационные объекты англоязычных резюме - это наборы слов, которые грамматически никак не согласованы (как и для русского языка). Их выделение может осуществляться по чисто формальным принципам. Например, адрес может рассматриваться как набор буквосочетаний 'P.O.', BOX, ST..., слов с большой буквы и чисел. Каждый такой набор может иметь свои границы и недопустимые компоненты. Например, в адресах не может быть ФИО, глаголов и т.д. Выделение таких наборов слов (описаний объектов) основано на использовании контекстных правил следующего вида:

CONTEXT(<слово1>,<слово2>,...,<словоN>) -> <результ. фрагмент>

где <слово1>,... это может быть - отдельное слово, признак, а также И-ИЛИ графы.

Для этих правил указывается, с какой позиции начинать применение, а также допустимый или недопустимый контекст. Далее, может быть указано, слово с какими признаками не должно стоять на той или другой позиции. Это обеспечивает дифференцированное применение правил. Все эти указания осуществляются с помощью фрагментов РСС.

Такие правила выделяют из текста группы слов (по их признакам), описывающих какой-либо объект, и заменяют их на одно слово, с которым связывается соответствующий фрагмент семантической сети, например, представляющий адрес.

Применение каждого правила - это последовательность действий, основанных на анализе слов и их признаков. Например, рассмотрим, как применяется правило, выделяющее словосочетания с предлогом OF.

Правило содержит специальный фрагмент, который указывает, что применять это правило нужно с 2-ой позиции, т.е. искать слова OF. Другой фрагмент отделяет левую часть от правой (->). В правой части стоит фрагмент, который указывает, что слова на 1-й и 3-й позициях должны быть склеены в комбинацию слов, которое в дальнейшем будет рассматриваться как одно слово с признаком OBJ. Это правило осуществляет преобразования:

СЛОВО с признаком объект OBJ или англ. СЛОВО (с признаком ENGL) + OF + СЛОВО с признаком объект OBJ или англ. СЛОВО -> <комбинация слов>

Контекстные правила применяются в строго определенной последовательности - каждое на своем уровне. Ниже приведен пример (см. пример 1) представления уровней англоязычной версии, определяющих порядок применения правил.

(1) Пример представления уровней
= Уровни =
LEVEL_ENG(LEVEL_1,LEVEL_2,LEVEL_3,LEVEL_4,LEVEL_5)
LEVEL_1(MORF_ENG) = Выявление частей речи англ. слов =
LEVEL_1(MORF) = Синонимы, термины =
LEVEL_2(TT~~1,TT~~2,TT~~3,TT~~4) = Выделение дат =
LEVEL_2(PP~~1,PP~~2) = Выделение мест - PLACE_ =
LEVEL_2(FF~~1,FF~~2,FF~~3,FF~~4) = Выявление лиц =
.....

Когда в соответствующих информационных объектах встречаются слова, то группы слов привязываются к той или иной позиции правила.

С помощью контекстных правил строится семантическая сеть - содержательный портрет документа (анкеты). Эта сеть с помощью обратного ЛП, который также управляется своими ЛЗ, отображается на поля сайта.

3. Интеграция системы в Интернет-портал

Программная несовместимость системы с порталом вызвана тем, что Интернет-портал работает на платформе Linux и реализован на языке Java. В то время, как ЛП реализован на языке Delphi для платформы MS Windows. Поэтому было принято решение реализовать ЛП в виде Веб-сервиса. Последний можно определить как приложение, предоставляющее некоторый сервис через Internet платформенно - и языково-независимым способом.

Данная реализация устранила проблемы, связанные с форматом обмена данными. Так как Веб-сервисы тесно связаны с XML, то проблема с передачей сложных типов данных (таких как составные типы данных и вложенные структуры) была решена описанием этих типов в формате XML.

Также была устранены экстралингвистические шумы. В рамках Интернет-портала текстовые данные хранятся и циркулируют в кодировке UTF-8, а при преобразовании в кодировку ср-1251, которая является основной кодировкой ЛП, часть символов корректно преобразуется их в соответствующие эквиваленты.

4. Входные и выходные параметры

Во входном параметре метода parse_cv поступает строка sCvPlainText в кодировке UTF-8 (см. пример 2).

(2) Пример входной информации
RESUME Ivanov Petr Andreevich home phone: 111 22 33
mobile: +903 111 22 33
Date of Birth: Feb. 01, 1981 e-mail: ivanov_p@yahoo.com
EXPERIENCE
1999 - 2005 Moscow Branch office of JVC. Assistant Assisted the Office Manager and Branch Office Representative. Processed documents and letters that supported office operations.
2002 - 2000 UNESCO (United Nations Educational, Scientific and Cultural Organization) in Kyrgyz Republic. Reception manager. Prepared operational documents and provided translation services. Worked with reports and statistic documents such as the "Annual report of Education of Youth".

В качестве результата работы системы ЛП, возвращается класс clsParsedCv в формате xml в UTF-8 (см. пример 3).
(3) Пример выходной информации
<CvXml>
<lastName>Andreevich</lastName>
<firstName>Ivanov</firstName>
<middleName>Petr</middleName>
<birthDay>01.02.1981</birthDay>
<email>ivanov_p@yahoo.com</email>
<homePhone>+7 (095) 111 22 33</homePhone>
<cellPhone>+7 (903) 111 22 33</cellPhone>
<workPhone />
<experienceEntry>
<StartDate>1999</StartDate>
<EndDate>2005</EndDate>
<Organization>Moscow Branch office of JVC.</Organization>
<Position>Assistant</Position>
<Description>Assisted the Office Manager and Branch Office Representative. Processed documents and letters that supported office operations.</Description>
</experienceEntry>
<experienceEntry>
<StartDate>2002</StartDate>
<EndDate>2000</EndDate>
<Organization>UNESCO (United Nations Educational, Scientific and Cultural Organization) in Kyrgyz Republic.</Organization>
<Position>Reception manager.</Position>
<Description>Prepared operational documents and provided translation services. Worked with reports and statistic documents such as the "Annual report of Education of Youth"</Description>
</experienceEntry>
</CvXml>

Заключение

Таким образом, разработанная система LINGVO-MASTER для англоязычных текстов решила задачу формализации текстов резюме на английском языке. Благодаря сетевой архитектуре она может быть использована большинством кадровых и рекрутинговых агентств вне зависимости от того, на какой платформе работает их программное обеспечение и на каком языке программирования оно реализовано.

Литература

1. *Кузнецов И.П., Мацкевич А.Г.* Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных. // Труды семинара Диалог-2006 по компьютерной лингвистике и ее приложениям (см. наст. сборник).