УДК 004.912 + 811.581

ИЗВЛЕЧЕНИЕ ИМЕН СОБСТВЕННЫХ ИЗ ТЕКСТОВ (NAMED ENTITY RECOGNITION) ПО ТЕМАТИКЕ ШОС НА КИТАЙСКОМ ЯЗЫКЕ ПРИ ПОМОЩИ ПРОСТЕЙШИХ СКРИПТОВ UNIX

Б.Г. Фаткулин

В статье рассматриваются методы прикладной лингвистики для извлечения имен собственных (named entity recognition) из китайских текстов по тематике ШОС. Объектом исследования является глава электронной энциклопедии Байду, посвященная ШОС. Подчеркивается необходимость сотрудничества между китайскими и российскими исследователями в области прикладной лингвистики, терминоведения, составления тезаурусов отдельных отраслей гуманитарных наук в Китае и России. Статья содержит пример практического использования простейших программных скриптов UNIX (grep, sed, tr) в редакторе vim для извлечения закавыченных имен собственных в китайских текстах. Приводятся примеры имен собственных и ссылка для скачивания полного списка.

Ключевые слова: Шанхайская организация сотрудничества, Китайская народная республика, терминология, прикладная лингвистика, извлечение имен собственных, синология, лингвистические корпуса, энциклопедия Байду, обработка естественных языков, междисциплинарные исследования, UNIX, программные скрипты, кавычки, named entity recognition, information extraction, information retrieval, natural language processing, table processing, table extraction, semi-structured information extraction, html, wiki markup.

В XXI в. одними из самых актуальных и обсуждаемых являются проблемы развития образовательных моделей, наращивания человеческого потенциала. Площадками, на которых поднимаются и решаются эти проблемы выступают такие международные и региональные объединения как Шанхайская Организация Сотрудничества (ШОС), Азиатско-Тихоокеанское Экономическое Сотрудничество (АТЭС), группа стран БРИКС, а также сетевые Университеты ШОС, АТЭС, БРИКС, Институты Конфуция.

Такой подход требует формирования профессионального сообщества и системы профессиональных коммуникаций, подготовки кадров специалистов-аналитиков, выработки основных квалификационных требований к профессии аналитика, поиска решений в области профессионального образования [2].

Профессия аналитика информационных ресурсов востребована во всех отраслях национальной экономики. Данная сфера профессиональной деятельности предполагает освоение широкого спектра знаний, умений и навыков, а также наличие определенного набора личных качеств. Количественные методы прикладной лингвистики могут использоваться аналитиками в текстологических исследованиях.

Необходимо отметить, что китайские специалисты по ШОС используют в своих работах собственную аутентичную терминологию, во многом отличную от терминологии глобальных англоязычных сетевых структур и ее эквивалентов в других европейских языках и русском языке. Источником терминологии выступают термины аутентичной китайской исторической науки, философия конфуцианства и теоретико-методологические установки КПК. Данные политические подходы Китайской народной республики широко представлены в программных документах Шанхайской организации сотрудничества, декларациях ШОС, документах, обслуживающих повседневную деятельность ШОС, научных статьях по тематике ШОС.

Named Entity Recognition (извлечение сущностей) – это одна из задач text mining [1], суть которой состоит в автоматическом определении сущностей в неструктурированных тестовых данных. Классическими сущностями выступают имена людей и компаний (names), адреса (locations), географические объекты (locations), даты (dates) и, в более сложных случаях, связи между ними, а также события, причинно-следственные связи, хронометраж событий. Также можно добавить к этому списку такие сущности, как электронные адреса, телефоны, определенные типы данных (например, IP адреса).

Таким образом, named entities составляют основу аналитики. В современной прикладной лингвистике существуют многочисленные методы извлечения терминологии из больших массивов текстов, называемых корпусами. Мы опробовали эти методы в процессе извлечения терминологии, когда использовали тексты главы «Шанхайская организация сотрудничества» в китайской электронной энциклопедии Байду. Baidu — интернет-энциклопедия на китайском языке, разрабатываемая и поддерживаемая Китайской поисковой системой Baidu. Также как и сама Baidu, энциклопедия подвержена цензуре в соответствии с правительственными постановлениями. На июнь 2013 года энциклопедия Байду содержит более 6,2 млн статей (больше, чем английская и немецкая Википедия вместе взятые); в ней зарегистрировано более 3,2 млн участников.

Полученные одиночные термины или терминологические словосочетания выводились из текста и переводились при помощи китайско-русского электронного переводчика с последующим ручным редактированием и поиском эквивалентов.

Особое внимание уделялось извлечению терминов из названий глав и подглав раздела «ШОС», представляющих информацию в виде информационной онтологии. В результате использования программного обеспечения тексты раздела «ШОС» были разбиты на смысловые группы. Обработанные сегментерами тексты получали синтаксическую и морфологическую разметку, границы слов четко обозначались. Полученные одиночные термины или терминологические словосочетания выводились из текста и переводились при помощи китайско-русского электронного переводчика с последующим ручным редактированием и поиском эквивалентов. Особое внимание уделялось извлечению терминов из названий глав и подглав раздела «ШОС», представляющих информацию в виде информационной онтологии.

Работа делилась на два этапа:

1) подготовка текста;

2) выделение «закавыченных» имен собственных с помощью алгоритмов Named Entity Recognition.

В основе нашего алгоритма лежит принцип пунктуации в китайском языке, согласно которому в китайском языке существуют специальные кавычки для выделения названий. Кавычки – парный знак препинания, который употребляется для выделения прямой речи, цитат, отсылок, названий литературных произведений, газет, журналов, предприятий, а также отдельных слов, если они включаются в текст не в своём обычном значении, используются в ироническом смысле, предлагаются впервые или, наоборот, как устаревшие и т.п.

Используя алгоритмы обработки естественных языков, мы открыли корпус текстов в редакторе vim и использовали скрипты, для того, чтобы перед каждым знаком открывающихся кавычек 《 вставить знак начала новой строки:

:% s/《 /\r《 /g

Затем после каждого знака закрывающихся кавычек 》поставили знак перевода каретки:

:% s/》 /\r》 /g

Одна из главных особенностей редактора – применение двух основных, вручную переключаемых, режимов ввода: командного (после запуска редактор находится в нём) и текстового (режим непосредственного редактирования текста, аналогичный большинству «обычных» редакторов). В командном режиме мы смогли набирать скрипты и исполнять их непосредственно в самом редакторе.

Таким образом, закавыченные выражение стали представлять собой отдельную строку. В результате применения вышеназванных методик исследования нами был составлен мультиязычный список названий ШОС на китайском языке с переводом на русский, английский, персидский и арабский языки, включающий 52 понятия.

Для иллюстрации результатов приведем примеры из этого списка:

《3 сань гу ши ли》&Три силы зла "&" Three evil forces"&" se niru-ye sheytany "&"qwa alshrr althlath "

5《 шан хэ цзучжи чжоунянь Асытаньа сюаньянь》&" 10-я годовщина Декларации ШОС Астане "&" 10th anniversary of the Declaration of the SCO Astana "&" aldkra 10 le'elan mnzmh shanghhay llt'eawn astana "&dahomin salgard eelan taasis sazman hamkary shanghay

Полный файл можно скачать и посмотреть по следующей ссылке: http://yadi.sk/d/LMjUUh1qLt9dr.

Библиографический список

1. Бацанина М.С. Информационный анализ лент деловых новостей / М.С. Бацанина // Труды СПбГУКИ. – 2013.

2. Логиновский, О.В. Информационно-аналитические центры как инструмент развития интеллектуального ресурса современного общества / О.В. Логиновский, В.Н. Любицын // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2012. – № 23.

3. Подчернин, В.М. Некоторые вопросы информационного обеспечения и информационно-аналитической деятельности / В.П. Подчернин // БИБЛИОСФЕРА. – 2007. – № 1.