УДК 004.912

С.Б. Данилевич

ХГУ «НУА», г. Харьков, Украина, danilevichsb@mail.ru

ПРИМЕНЕНИЕ МЕТОДОВ КОРПУСНОЙ ЛИНГВИСТИКИ ДЛЯ ПОЛУЧЕНИЯ ИНФОРМАЦИИ НА НЕЗНАКОМОМ ЯЗЫКЕ

Данная работа посвящена использованию доступного инструментария корпусной лингвистики для выявления информации об интересуемом объекте в текстах, представленных в Интернет на незнакомом пользователю языке.

КОРПУСНАЯ ЛИНГВИСТИКА, ЛИНГВИСТИЧЕСКАЯ ПРАГМАТИКА, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, АНАЛИЗ ТЕКСТА, ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ

Введение

Деловые, профессиональные и культурные связи Украины с другими странами постоянно укрепляются, что требует от выпускников любой специальности постоянного расширения кругозора, в частности, знакомства с методами корпусной лингвистики, которые могут быть полезны для анализа текстов на незнакомом языке.

Лингвистический корпус (linguistic corpus, text corpus) – это собрание отрывков текстов в электронной форме, отобранных в соответствии с внешними критериями, чтобы наиболее полно представлять язык или вариацию языка. Функционирует как источник данных для лингвистических исследований (John Sinclair) [1].

Для английского (http://www.natcorp.ox.ac.uk), немецкого (http://corpora.ids-mannheim.de/), чешского (http://ucnk.ff.cuni.cz), японского (http://cblle.tufs.ac.jp), русского (http://ruscorpora) и др. языков созданы полноценные корпусы, соответствующие основным требованиям.

Интенсивно разрабатывается корпус текстов украинского языка сотрудниками лаборатории компьютерной лингвистики Института филологии Киевского национального университета имени Тараса Шевченко под руководством Н.П. Дарчук (http://www.mova.info).

Привлечение методов корпусной лингвистики активно используется в обучении иностранным языкам, переводоведении, преподавании перевода, для повышения эффективности аналитической обработки научной информации, представленной в виде корпуса распределенных текстовых документов, расположенных на различных Web-ресурсах.

Специальное программное обеспечение дает возможность искать в корпусе необходимую информацию. Программы частотного анализа текстов используются для проведения информометричных исследований в информационных ресурсах как предпосылки выделения из хранилищ данных новых знаний. Частотный анализ, например, материалов конференций за несколько лет [2], позволяет выявить новые знания и тенденции развития исследуемого направления, экономить время, отбирая из больших массивов только те статьи, которые содержат полезную информацию. При этом для оперативного составления частотного словаря пригодны стандартные приложения MS Office [3]. Однако формат частотного словаря не позволяет решить проблемы снятия полисемии, омонимии и других видов неоднозначности.

Наиболее информативным является поиск контекстов слов. Использование информационных технологий сделали возможной обработку больших массивов текстов, в частности, представленных в Интернет, который в свою очередь можно считать огромным текстовым корпусом. Значительная часть такой информации хранится в виде неструктурированного текста, для анализа которого разработано достаточно много программных продуктов. Профессионалами применяются методы Text Mining и Data Mining, реализованные в программах: WordStat, TextAnalyst, Businessobjects Text Analysis, AeroText, STATISTICA Text Miner, Attensity suite, Weka, Galaktika-ZOOM («Галактика»), Медиалогия, WordSmith Tools. Качественная Data Mining-программа может стоить достаточно дорого, быть относительно сложной и узконаправленной [4]. В то же время, извлечение полезных сведений невозможно без понимания сути данных, поэтому применение методов Data Mining не может заменить аналитика [4].

Для поиска, управления данными в корпусе, их статистической обработки, предоставления результатов в удобной форме применяются программные средства – корпус-менеджеры (corpus manager). Одной из наиболее удачных программ анализа корпусов для выявления особенностей употребления слов является платная и работающая только под MS Windows утилита, разработанная в Оксфордском университете WordSmith Tools (http://www.lexically.net/wordsmith). Существуют свободно распространяемые корпус-менеджеры. Так, в Германии разрабатывается конкордансер Corsis (http://corsis.sourceforge.net). AntConc (http://www.antlab.sci.waseda.ac.jp/software.html) не требует установки, работает под MS Windows,

Linux и Mac, распознает свыше 90 кодировок. Есть ограничения на формат входных файлов (htm, html, xml, txt). Результат сохраняется в txt-формате. Использование таких программ кроме лингвистов может быть полезно самым разным исследователям – политологам, социологам и др. Так, например, автоматическое создание и последующий анализ частотных словарей и конкордансов, полученных на сформированном корпусе текстов статей, представленных в электронном виде в сети Интернет был использован находящаяся в свободном доступе корпус-менеджер TextSTAT для изучения способов создания имиджа страны в текстах периодической печати [5].

В данной статье рассматривается работа с программами TextSTAT и AntConc. К сожалению, эти программы не поддерживают кириллицу.

Основной ***целью статьи*** является оценка возможности использования методов корпусной лингвистики для выявления информации об интересуемом объекте в текстах из Интернет на незнакомом языке, (например, чешском) с помощью доступного бесплатного программного обеспечения и приложений MS Office.

1. Построение текстового корпуса

Поверхностный анализ текстов из Интернета на незнакомом пользователю языке, сравнение полученной информации с аналогичной информацией с отечественных сайтов можно производить при помощи создания соответствующих текстовых корпусов.

Построить свой корпус позволяет свободно распространяемая программа TextSTAT, бесплатно её можно скачать, например, по адресу: http://www.softpedia.com/get/Office-tools/Other-Office-Tools/TextStat.shtml. Программа поставляется в виде отдельного файла установки. Чтобы установить программу, достаточно распаковать файл в папку по выбору. После запуска программы открывается диалоговое окно для создания собственного текстового корпуса.

В качестве тематики исследования можно выбрать, например, проблемы в высшем образовании в Чешской республике для сравнения с аналогичными отечественными проблемами. Для получения данных для анализа целесообразно использовать Web-страницы, полученные по запросу в чешском поисковике seznam.cz: «problémy v oblasti vysokoškolského vzdělávání (проблемы в высшем образовании)».

Далее нужно добавить на свой корпус соответствующие тематике файлы из Интернета. Программа TextSTAT читает неограниченное число страниц с выбранного сайта на английском, немецком, голландском, португальском, испанском, галисийском (официальном языке испанской Галисии), французском, итальянском, финском, польском, чешском языках и переносит их в программу, убирая HTML-теги. Для добавления файла используется окно Web-паука (Web Spider), (рис. 1), в котором вводится соответствующий Web-адрес. На массиве данных текстов, полученных из соответствующих Web-сайтов могут быть созданы:

1. корпус текстов данной тематики (файл с расширением .crp) – вкладка Korpus,
2. частотный словарь – вкладка Tvary slov,
3. конкорданс – вкладка Konkordance.

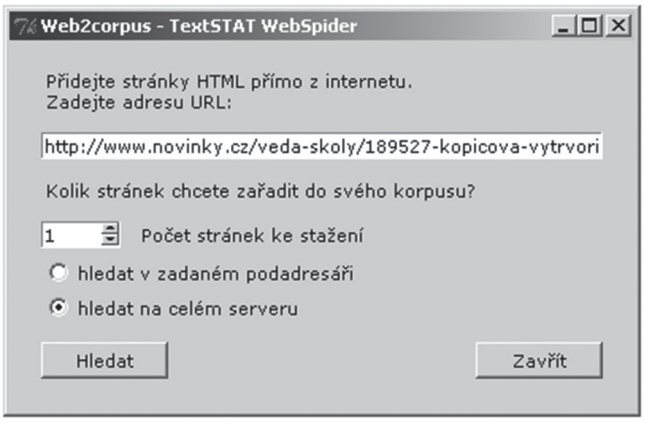

Рис. 1. Web-паук (Web Spider) программы TextSTAT

Для примера выбраны первые 10 ссылок по данному запросу (рис. 2).

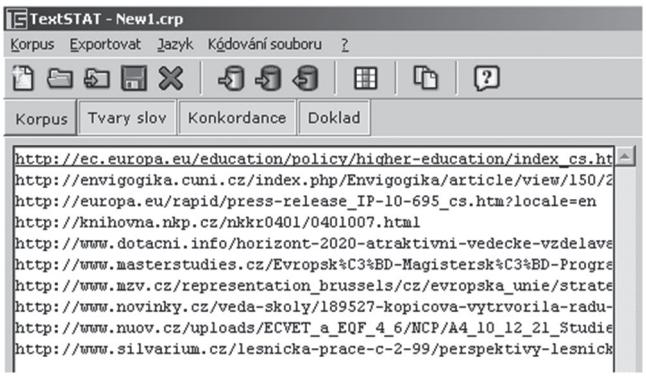

Рис. 2. Добавление файлов из сети Интернет

Программа извлекает найденные слова в список, которые, после этого, могут быть отредактированы вручную. Есть возможность подключить так называемый список «стоп-слов» - список из слов, которые не относятся к терминологии. На вкладке Конкорданс, введя интересующее слово, можно получить это слово в контексте данного корпуса и перейти для более детального анализа на Web-страницу, содержащую это слово в данном контексте. При поиске можно использовать специальные символы, используемые в регулярных выражениях:

«.» – (точка) обозначает любой символ;

«\ w» – шаблон для любых буквенно-цифровых символов;

«\ W» – шаблон для любых неалфавитно-цифровых символов (например, пробел, знаки препинания);

«+» – сочетание с предыдущим символом, которое повторяется один раз или несколько раз (например, ab + c соответствует «abc», «abbc», «abbbc» и так далее, но не «ac»);

«*» – сочетание с предыдущими символами повторяется любое число раз, включая нуль;

«|» – обозначает «или»;

«[]» – в квадратных скобках определяет набор символов, которые ищутся альтернативно.

2. Анализ текстового корпуса

Анализ полученного текстового корпуса удобно производить с помощью программы AntConc (http://www.laurenceanthony.net/antconc_index.html). Корпус-менеджер AntConc (Dr. Laurence Anthony) используется для получения словарных минимумов, списков устойчивых сочетаний, выборок к тематическим группам слов [3].

Чтобы воспользоваться полученной и очищенной от тэгов информацией TextSTAT в AntConc, нужно сохранить корпус, переименовав файл с расширением .crp в txt-файл. После чего данный корпус можно открыть в AntConc.

Для поиска и извлечения информации из корпуса текстов можно воспользоваться следующей процедурой: создать частотный словарь, проанализировать его, отобрать значимые слова, составить конкорданс и исследовать контекст употребления значимых слов.

Частотный словарь корпуса создается на вкладке Word List, что уже позволяет получить большое количество информации об употреблении слов в корпусе. Этот словарь можно после сохранения в текстовом формате (File – Save Output to text File...) открыть в Excel (Данные – Получить внешние данные – Из текста – antconc_results.txt – Импорт). В окне Мастера импорта Excel следует указать формат данных: с разделителем, Формат файла: 65001: Юникод UTF-8. На втором шаге Мастера импорта необходимо установить Символом-разделителем знак табуляции. На третьем шаге Мастера нужно указать Формат данных столбцов: текстовый. В результате получится частотный словарь в офисной программе Excel, что позволяет воспользоваться всеми возможностями MS Office, включая офисное программирование.

Дело в том, что наряду со значимыми ключевыми словами в этот перечень попадают элементы не несущие смысловой нагрузки (например, одно- и двухбуквенные словоформы, аббревиатуры и т.п.). Здесь можно воспользоваться фильтрацией Excel (рис. 3).

Отфильтровав одно-, двухбуквенные и т.д. слова с помощью шаблона «?» – один любой знак, дав команду на вкладке Главная: Найти и выделить – Перейти – Выделить – Только видимые ячейки можно избавиться от некоторых незначащих слов (рис. 4).

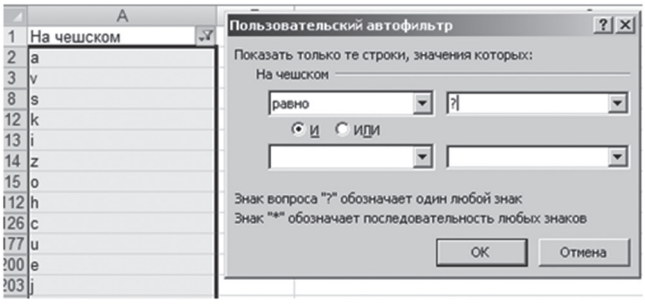

Рис. 3. Фильтрация однобуквенных слов

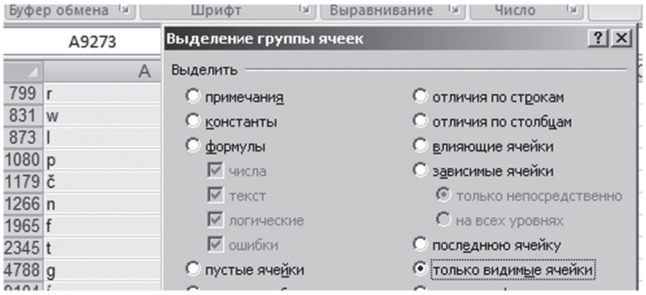

Рис. 4. Выделение только однобуквенных слов для их удаления

Более тщательная подготовка текста требует удаления и других слов («стоп-слова»). Помощь здесь могут оказать готовые перечни стоп-слов. Так, для чешского языка такой список можно найти по адресу: http://nlp.fi.muni.cz/cs/StopList.

Вычитание стоп-списка из полученного частотного словаря можно произвести в Excel с помощью расширенного фильтра или используя VBA. Аналогично используется словарь синонимов (например, ABZ slovník českých synonym - on-line hledání – http://www.slovnik-synonym.cz/). Проблема остаётся с собственными именами, ошибками в самом тексте, употреблением слов на других языках и т.п.

Перевод слов частотного словаря можно получить автоматически, воспользовавшись функцией перевода с одного языка на другой (с использованием Google Translate), написанной на языке VBA (http://excelvba.ru/code/GoogleTranslate) (рис. 5).

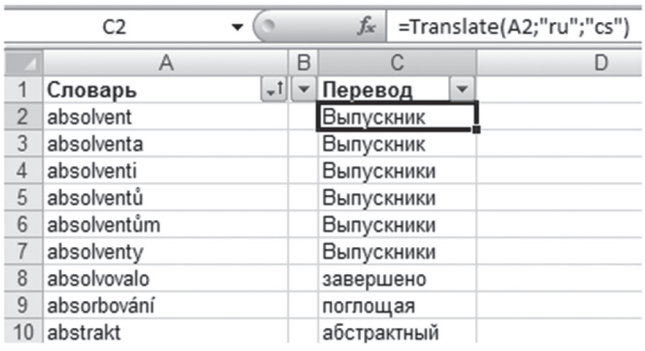

Рис. 5. Автоматический перевод

Одним из существенных недостатков является то, что каждая морфологическая форма одного и того же слова подсчитывается как отдельное слово. Тем не менее, полученный список позволяет отобрать наиболее значимые с точки зрения исследователя ключевые слова для поиска контекста.

3. Извлечение информации из конкорданса

Для определения контекста использования выбранного слова можно использовать в AntConc функцию Concordance (список контекстов ключевого слова в исследуемом корпусе). AntConc обеспечивает показ списка всех словоупотреблений анализируемого текста с ближайшим контекстом нужного количества слов (обычно 5–6) перед и после искомого термина, список n-грамм (последовательностей из n слов) любой длины и частоты для данного корпуса.

Например, выбрав в качестве ключевого слова – priority, можно получить всю фразу с этим словом (рис. 6), а при необходимости и весь текст. Результаты поиска показываются в формате KWIC (key word in context).

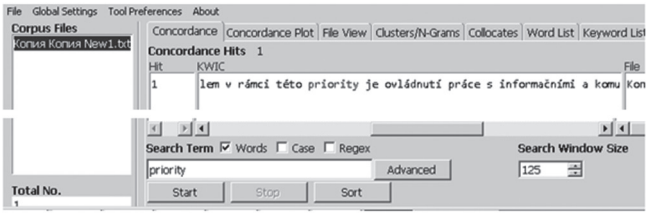

Рис. 6. Функция Concordance AntConc

Так, фраза: «Základním cílem v rámci této priority je ovládnutí práce s informačními a komunikačními technologiemi studenty i učiteli», а её Google-перевод: «Основная цель данного приоритета является овладение работой со студентами и преподавателями в области ИКТ».

Таким образом, последовательность действий для определения проблем, которые видят в Чешской республике в области высшего образования следующая:

– установка программ TextSTAT, AntConc;

– выбор Web-страниц по данной тематике;

– поставка этих страниц в TextSTAT и сохранение, полученной очи-щенной от тэгов информации в формате txt;

– в AntConc открытие текстового файла, созданного в TextSTAT;

– создание частотного словаря (Word List), который может быть открыт в Excel;

– перевод слов частотного словаря в переводчике Google;

– использование в AntConc функции Concordance для определения контекста использования выбранного слова;

– перевод контекста.

Сравнение проблем в высшем образовании в Украине и Чехии, упомянутых на данных сайтах в данный момент времени, тем не менее, позволяет сделать некоторые выводы. Так, часть проблем совпадают: недостаточное финансирование; трудности выпускников в поиске работы и в то же время проблемы у работодателей с нехваткой персонала, которые отвечают потребностям современного рынка труда; проблемы, связанные с естественно-математической подготовкой и др. На данных украинских сайтах были упомянуты специфические проблемы, такие как, проблемы со стандартами, существования высших учебных заведений разных форм собственности, внедрение реформ, организация учебно-воспитательной работы и др. На данных чешских сайтах упоминались проблемы, которых не было на данных украинских. Например, проблемы с содействием социального и культурного разнообразия. Много внимания было уделено психологическим, поведенческим проблемам, вызывающие трудности в обучении особенно у людей с ограниченными физическими возможностями. Отмечалось недостаточное инвестирование в высшее образование (в ЕС счетов лишь 1,3% от ВВП по сравнению с 3,1% в США и 1,5% в Японии) и надежда, что государственное финансирование будет дополнено соответствующими дополнительными ресурсами, с большим участием частного сектора. Не определены компетенции выпускника ВУЗа, неэффективны механизмы стимулирования и отсутствие системного развитие непрерывного образования взрослых Решение некоторых проблем видят в увеличении квалификации и профессионального мастерства преподавателей, разработке и внедрении системы финансовых и нефинансовых стимулов для работодателей увеличить расходы на обучение персонала и др.

Выводы

Представленная методика и практические рекомендации по применению методов корпусной лингвистики, использованию доступных программных средств (TextSTAT, AntConc), стандартных приложений MS Office, элементов офисного программирования позволяют в некоторой степени автоматизировать поиск нужной информации из сети Интернет и предварительный анализ текстов на незнакомом языке. Приводится конкретный пример поиска информации на чешских сайтах, касающейся проблем высшей школы. При достаточной заинтересованности можно составить корпус большего объема, что позволит глубже анализировать информацию. Знакомство с основами корпусной лингвистики полезно для представителей любой специальности.

Список литературы: 1. *Sinclair, J.* Corpus, Concordance, Collocation. – Oxford University Press, 1991. – 197 p. **2.** *Кузнецов А.Ю.* Информометрические исследования докладов конференций Крым / А.Ю. Кузнецов // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса [Электронный ресурс]: 18-я Междунар. конф. "Крым 2011": Тр. конф. — Электрон. дан. — М., 2011. — http://www.gpntb.ru/win/inter-events/crimea2011/disk/042.pdf. **3.** *Данилевич С.Б.* Создание частотных словарей доступными средствами в учебных целях // Обучение иностранных студентов в высшей школе: традиции и перспективы: Материалы Международной научно-методической конференции (23-24 мая 2013 г., Харьков) [Текст]: тезиси. — Харьков: НТУ «ХПИ», 2013. — 264 с. — С. 121–123. **4.** *Чубукова И.А.* Data Mining: учебное пособие. — М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. — 382 с. **5.** *Зверева П.П.,* Максименко О.И. Современные направления лингвистических исследований имиджа страны и её жителей. [Текст]: статья. — Вестник МГОУ. Серия «Лингвистика». – № 6, 2013. – С. 25-29.

Поступила в редколлегию 25.09.2014

УДК 004.912

Застосування методів корпусної лінгвістики для отримання інформації на незнайомій мові / С.Б. Данилевич // Біоніка інтелекту: наук.-техн. журнал. – 2014. – № 2 (83). – С. 19–23.

У статті розглядаються методика і практичні рекомендації щодо застосування методів корпусної лінгвістики, використанню доступних програмних засобів, елементів офісного програмування для автоматизації пошуку потрібної інформації з мережі Інтернет і попереднього аналізу текстів незнайомою мовою.

Іл. 6. Бібліогр .: 5 найм.

UDK 004.912

Application of Corpus Linguistics Methods to Acquire Information in a Foreign Language / S.B. Danylevych // Bionics of Intelligence: Sci. Mag. – 2014. – № 2 (83). – P. 19–23.

This paper presents methods and practical recommendations for corpus linguistics methods, available software and elements of the office programming to be applied for automation of the process of the required information retrieval in the Internet and preliminary analysis of the source texts.

Fig. 6. Ref.: 5 items.