КУЛЬТУРОЛОГИЯ, ФИЛОЛОГИЯ, ИСКУССТВОВЕДЕНИЕ: АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОЙ НАУКИ

Сборник статей по материалам
XXII международной научно-практической конференции

№ 5 (17)
Май 2019 г.

Издается с августа 2017 года

Новосибирск
2019

СЕКЦИЯ

«ПРИКЛАДНАЯ И МАТЕМАТИЧЕСКАЯ ЛИНГВИСТИКА»

ВНЕДРЕНИЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ В КОРПУСНУЮ ЛИНГВИСТИКУ

Абдуллаев Сайяр Габиб оглы
заведующий отделением
Институт Информационных Технологий
Национальной Академии Наук Азербайджана (ИИТ НАНА),
Республика Азербайджан, г. Баку
E-mail: depart5@iit.ab.az

Абасова Судаба Ейбалы гызы
ст. научный сотрудник
Институт Информационных Технологий
Национальной Академии Наук Азербайджана (ИИТ НАНА),
Республика Азербайджан, г. Баку

АННОТАЦИЯ

В работе показана значимость аппаратного, системного и прикладного программного обеспечения информационных технологий, автоматической обработки лингвистических данных и внедрения рабочего места лингвиста. Представлена информация о различных аспектах корпусной лингвистики и их правилах, были выдвинуты на первый план преимущества создания национальных корпусов.

Ключевые слова: лингвистика, лингвистический корпус, корпусное языкознание, национальный корпус.

Введение. С целью проведения больших вычислений над лингвистическими данными, в том числе и лингвистического моделирования наиболее приемлимым является использование вычислительных машин (либо компьютеров). Для выполнения какого-либо действия при помощи компьютера наряду с апаратным обеспечением нужен (hardware) набор команд – программ. Программное обеспечение компьютера (software),

являясь неотъемлемой частью компьютерной системы, является логическим продолжением технического обеспечения компьютера. Конкретная сфера применения компьютера определяется его программным обеспечением. Аппаратное (hardware) и программное (software) обеспечение информационных технологий тесно связаны друг с другом [1].

Программное обеспечение (ПО) – это компьютерные программы, написанные на основе последовательных команд на машинном языке, для управления аппаратными средствами и выполнения различных операций над информацией, а также для соответствующей документации. Системное и прикладное программное обеспечение отличаются друг от друга в зависимости от назначения программных средств. Системные программы служат для управления работой аппаратных средств и загрузки операционных систем, утилит, драйверов и ряда других программ. Прикладная программа предназначается для конечного пользователя и помогает ему производить различные операции над информацией: позволяет создавать и обрабатывать тексты (текстовые редакторы), графические изображения (графические редакторы), работать над звуковой и видео информаций (мультемедийные программы), создавать электронные таблицы (электронные таблицы) для обработки статистических данных) и т. д.

Специально для лингвистики создаются такие прикладные программы как электронный перевод и словари, в том числе и мультимедийные учебные программы. Некоторые исследователи наряду с аппаратным и программным обеспечением пользуются понятием linqvare (или linquvare) которое обобщает все лингвистические ресурсы (грамматические данные, словари, энциклопедии, лингвистическая база данных и т. д.). Важные для автоматической обработки лингвистических данных аппаратные, программные и лингвистические средства в совокупности называются Автоматическим Рабочим Местом (АРМ) лингвиста. Родной язык и внедряющийся иностранный язык, в том числе различные лингвистические компьютерные ресурсы, операционная и прикладное базовое программное обеспечение (ПО) и компьютер являются составными частями АРМ. В зависимости от особенностей АРМ лингвиста можно усовершенствовать при помощи прикладных программ и лингвистических ресурсов изучения иностранного языка и перевода [2, 3].

Корпусная Лингвистика. Одной из важных задач лингвистики является сбор и хранение источников материалов для лингвистических исследований. На данный момент для решения таких задач пользуются набором текстов больших объемов хранение которых более выгодно в электронном виде. Использование компьютеров и специальных телекоммуникационных сетей подходит не только для хранения в электронном

виде текстов большого объема, но и дает возможность проводить по ним поиск, обрабатывать их и т.д. Задача сбора текстов или корпусов в электронном виде настолько важна для современной лингвистики, что сбор этих электронных текстов стал объектом исследования особого раздела прикладной лингвистики – корпусной лингвистики. Корпусная лингвистика, являясь разделом языкознания, занимается разработкой общих принципов построения и использования лингвистических корпусов при помощи компьютера. Таким образом, корпусную лингвистику можно выявить двумя ниже приведенными аспектами [4, 5]:

- Создание текстовых путем использования автоматических инструментов;
- Разработка способов исследования различных уровней языка на базе различного типа корпусов.

Современные исследователи-языковеды могут вести исследования на основе корпусов (доступных для общего пользования) созданых ими самими или другими исследователями и их коллективами. Помимо научных исследований корпусами пользуются в нижеследующих случаях:

- При создании лингвистических словарей, определении многозначных слов и т. д.;
- При определении частотности морфем в грамматике, типа словосочетаний, предложений и т. д.;
- Для определения связи между абзацами или внутри абзацев с целью различения типов лингвистических текстов друг от друга и т.д.;
- При автоматическом переводе текстов с целью поиска контекста слова имеющего ряд переводческих эквивалентов, при поиске эквивалентных переводов в параллельных текстах и т. д.;
- Поиск цитат, фрагментов произведений с целью обучения, примеры для организации учебных занятий, создание учебных средств и т. д.;
- При тестировании программ автоматического анализа и синтеза речи и т. д. [6, 7, 8].

Основное понятие корпусной лингвистики – лингвистический корпус определяется как набор специально выбранных текстов размеченных различными лингвистическими параметрами и обеспеченных поисковой системой. Таким образом, корпус можно охарактеризовать как Корпус = Тексты + Их разметка. В более широком смысле, это любая совокупность текстов. По этому признаку различают размеченные либо неразмеченные текстовые корпуса. Примером таких неразмеченных корпусов могут служить уже существующие наборы электронных текстов: виртуальные библиотеки, архивы электронных версий периодических изданий или новостных лент, которые оказываются достаточными для некоторых исследовательских и учебных целей.

Использование неразмеченных текстовых наборов содержащих поисковые инструменты увеличивает объем информации, и данная информация наряду со своей нерелевантностью создает трудности для пользователя при ее использовании. В связи с этим размеченные корпусы можно считать предметом корпусной лингвистики. На первом этапе создание корпуса начинается с выбора текстов. В этом случае нужно думать о том тексты, какого функционального стиля и конкретного жанра, какого года издания и в каком количестве будут добавлены в корпус.

При отборе текстов для создания корпуса нужно обратить внимание к нижеследующим требованиям:

- Репрезентативность (частота явления в корпуса должна совпадать с его частотой в естественном языке);
- Полнота (даже если явление неподходит идее репрезентативности оно должно быть включено в корпус);
- Достаточный объем (если объем первоначальных корпусов исчислялся миллионами слов, то сейчас их объем исчисляется сотнями миллионов и миллиардами; например, объем корпуса английского языка Bank of English составляет 2,5 миллиарда слов);
- Экономичность (при исследовании проблемной области корпуса текстов изначально должны создаваться таким образом, чтобы они экономили время исследователя, т. е. быть не только подмножеством текстов проблемной области, но и по возможности они должны быть «экономичными»);
- Структуризация материалов (в корпусе должны быть указаны адекватные ему единицы измерения);
- Компьютерная поддержка (поддержание корпуса текстов комплексом программой по обработки данных, при помощи которых можно определить контекст слов, статистическую инвентаризацию, автоматическую обработку информации и т. д.).

Самым главным этапом создания корпуса является его разметка. Разметка (англ. tagging, annotation) – это присвоение тексту и его компонентам специальной метки. Эти метки могут быть как внешними (экстралингвистическими) содержащими информацию об авторе и тексте, так и внутренними. Внутренние же в свою очередь, могут быть структурными или лингвистическими. Внешняя разметка дает информацию об авторе, о названии текста, о месте и годе, а также жанре издания. Информация об авторе может состоять не только об имени, но и об его возрасте, поле, годах жизни и т. д. Такая кодировка информация называется мета разметка. Структурная разметка дает информацию о статусе каждой единицы (глава, абзац, предложение, форма слова), а специальная лингвистика отражает лексические, грамматические и другие элементы текста. В соответсвии с уровнем лингвистического уровня различают

морфологические (определение морфологической категории и части речи), синтаксические (определение синтаксических связей), семантические (категории, характеризующие семантику слов), анафорические (характеристика референтных связей, например, местоимений), просодические (характеристика ударения и интонации), дискурсные (определение пауз, повторов, исправление устной речи) и другие виды разметок [9, 10].

В зависимости от особенностей сбора корпусов, их разметки, а также других факторов различают виды корпусов. Среди корпусов, созданных для различных национальных языков, самым важным является универсальный национальный корпус. Создание и расширение универсальных национальных корпусов является одной из главных задач корпусной лингвистики. Универсальный национальный корпус – это совокупность определенных текстов собранных на каком-либо естественном языке, с целью исследования этого естественного языка [11]. Британский национальный корпус (BNC) (www.natcorp.ox.ac.uk) является универсальным корпусом, который принят всеми [12]. Таким представительным корпусом для русского языка является Национальный корпус русского языка (НКРЯ) (www.ruscorpora.ru) [13]. Среди корпусов славянских языков выделяют Чешский Национальный Корпус (http://ucnk.ff.cuni.cz), созданный в Карловом Университете Праги [14]. Национальные корпусы также существуют и для японского, финского, немецкого и других языков. Одним из первых известных корпусов является Брауновский корпус, созданный в 1963 году в Брауновском Университете (США) с целью использования американского варианта словаря английского языка. Его объем составлял 1 миллион слов. Корпус был создан У. Френсисом и Г. Кучером. Ими была разработана строгая процедура по отбору текстов. В этот корпус были включены 500 фрагментов прозаических текстов американских авторов и в 1961 году, используя для каждого текста 2000 слов, он был опубликован. Тексты представляли собой 15 наиболее распространенных жанров художественной и информативной прозы [15]. В соответствии с запросом пользователя поиск в корпусе ведется на основе специальных программ называемых корпусными менеджерами. Корпусный менеджер (англ. corpus manager) – это специальная поисковая система, которая обеспечивает предоставление информации в удобной форме для пользователя и включающая в себя программы средства поиска данных в корпусе. Обычно результаты поиска выдаются в форме конкорданса (корпусных менеджеров также называют конкордансерами), где искомая единица представляется со своим контекстным окружением с частотными характеристиками граммем, языковых единиц и т. д. Таким образом, корпус представляет собой совокупность размеченных текстов объемом слов не менее 100 млн. слов, и дает возможности для прикладных и исследовательских целей.

Задачи компьютерной лингвистики. Компьютерная лингвистика возникла на стыке таких наук, как лингвистика, математика, информатика (Computer Science) и искусственный интеллект. Истоки КЛ восходят к исследованиям известного американского ученого Н. Хомского в области формализации структуры естественного языка ее развитие опирается на результаты в области общей лингвистики (языкознания). Языкознание изучает общие законы естественного языка, его структуру и функционирование, и включает такие области: Фонология – изучает звуки речи и правила их соединения при формировании речи; Морфология – занимается внутренней структурой и внешней формой слов речи, включая части речи и их категории; Синтаксис – изучает структуру предложений, правила сочетаемости и порядка следования слов в предложении, а также общие его свойства как единицы языка (ЕЯ). Семантика и прагматика – тесно связанные области: семантика занимается смыслом слов, предложений и других единиц речи, а прагматика – особенностями выражения этого смысла в связи с конкретными целями общения; Лексикография описывает лексикон конкретного ЕЯ – его отдельные слова и их грамматические свойства, а также методы создания словарей. Результаты Н. Хомского, полученные на стыке лингвистики и математики, заложили основу для теории формальных языков и грамматик (часто называемых генеративными, или порождающими грамматиками). Эта теория относится ныне к математической лингвистике и применяется для обработки не столько ЕЯ, но искусственных языков, в первую очередь – языков программирования. По своему характеру это вполне математическая дисциплина. К математической лингвистике относят также и квантитативную лингвистику, изучающую частотные характеристики языка – слов, их комбинаций, синтаксических конструкций и т. п. При этом используется математические методы статистики, так что можно назвать этот раздел науки статистической лингвистикой. КЛ тесно связана и с такой междисциплинарной научной областью, как искусственный интеллект (ИИ), в рамках которого разрабатываются компьютерные модели отдельных интеллектуальных функций. Одна из первых работающих программ в области ИИ и КЛ – это известная программа Т. Винограда, которая понимала простейшие приказы человека по изменению мира кубиков, сформулированные на ограниченном подмножестве ЕЯ. Отметим, что несмотря на очевидное пересечение исследований в области КЛ и ИИ (поскольку владение языком относится к интеллектуальным функциям), ИИ не поглощает всю КЛ, поскольку она имеет свой теоретический базис и методологию. Общим для указанных наук является компьютерное моделирование как основной метод и итоговая цель исследований. Таким образом, задача КЛ может быть сформулирована

как разработка компьютерных программ для автоматической обработки текстов на ЕЯ. И хотя при этом обработка понимается достаточно широко, далеко не все виды обработки могут быть названы лингвистическими, а соответствующие процессоры – лингвистическими. Лингвистический процессор должен использовать ту или иную формальную модель языка (пусть даже очень простую), а значит, быть так или иначе языково-зависимым (т. е. зависеть от конкретного ЕЯ). Так, например, текстовый редактор Microsoft Word может быть назван лингвистическим (хотя бы потому, что использует словари), а редактор NotePad – нет. Сложность задач КЛ связана с тем, что ЕЯ – сложная многоуровневая система знаков, возникшая для обмена информацией между людьми, выработанная в процессе практической деятельности человека, и постоянно изменяющаяся в связи с этой деятельностью. Другая сложность разработки методов КЛ (и сложность изучения ЕЯ в рамках языкознания) связана с многообразием естественных языков, существенными отличиями их лексики, морфологии, синтаксиса, разные языки предоставляют разные способы выражения одного и того же смысла [16].

Заключение. Кроме вышеперечисленных средств (анализ и синтез устной речи, автоматический ввод текстов, автоматическая обработка текста, использование текстовых корпусов, компьютерное обучение языку и т. д.) внедрения информационных технологий в лингвистику, информатика и лингвистика пересекаются в следующих сферах: получение знаний из текста, автоматическое индексирование и разделение на части документов, гипертекстовые технологии в лингвистике и т. д.

Список литературы:

1. Надежда Аппаратное и программное обеспечение компьютера - единое целое,http://www.compgramotnost.ru/vvedenie/chto-takoe-computer, 2010.
2. А.Н. Степанов, Информатика: учеб. пособие. СПб.: Питер, 2006.
3. Л.Ю. Щипицина, Информационные технологии в лингвистике: учеб. пособие, Л.Ю. Щипицина, М.: ФЛИНТА: Наука, 2013.
4. Г.Г. Белоногов, Компьютерная лингвистика и перспективные информационные технологии. М.: Русский мир, 2004.
5. А.В. Зубов, И.И. Зубова, Информационные технологии в лингвистике: учеб. пособие для студ. вузов. М.: Академия, 2004.
6. Е.М. Чухарев, Компьютерные технологии в лингвистических исследованиях: указания по выполнению домашнего задания. Архангельск, 2009.
7. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.
8. Б.И. Большакова, Компьютерная лингвистика: методы, ресурсы, приложения, Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011.

9. https://ru.wikipedia.org/wiki/Корпусная_лингвистика.
10. А.В. Всеволодова, Компьютерная обработка лингвистических данных: учеб. пособие. 2-е изд., испр. М.: Флинта: Наука, 2007.
11. https://ru.wikipedia.org/wiki Национальный_корпус.
12. https://ru.wikipedia.org/wiki/Британский_национальный_корпус.
13. www.ruscorpora.ru, Национальный корпус русского языка.
14. https://ru.wikipedia.org/wiki/Чешский_национальный_корпус.
15. http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html, Браун корпус.
16. Анечка Тимошко Компьютерная лингвистика: методы, ресурсы, приложения, https://pandia.ru/text/80/678/74827.php.

ЭФФЕКТИВНОСТЬ ИСПОЛЬЗОВАНИЯ РОДНОГО ЯЗЫКА В РАБОТЕ НАД ТЕКСТОМ

Магзумова Алма Таужановна
ст. преподаватель
Кокшетауского государственного университета
им. Ш.Ш. Уалиханова,
Республика Казахстан, г. Кокшетау
E-mail: magzumova.61@mail.ru

EFFICIENCY OF USE OF THE NATIVE LANGUAGE IN WORKING WITH TEXT

Alma Magzumova
senior Lecturer of the Kokshetau State University
named after Sh.Sh. Ualikhanov,
Kazakhstan, Kokshetau

АННОТАЦИЯ

Цель статьи заключается в рассмотрении эффективности использования родного языка в работе над текстом.

Особое внимание уделяется словарной работе,которая занимает особое место в процессе речевой деятельности студентов на первоначальном уровне. А также были рассмотрены методические и педагогические подходы исследователей. Предложенные в статье виды работ