ЗНАЧЕНИЕ ТЕКСТОВЫХ БАЗ ИНТЕРНЕТА В ЭФФЕКТИВНОМ РАЗВИТИИ ЯЗЫКОВОГО КОРПУСА

Норпулотова Мунира Шомурод кизи
Ташкентский государственный университет узбекского языка и литературы имени Алишера Навои. переподавать

Для создания корпуса удобно использовать Интернет. Потому что если письменные тексты определенного языка находятся в электронном варианте или в виде книги в корпусе, они сканируются и переводятся в словесную (словарную) форму, и в тексте встречается множество ошибок, которые необходимо исправлять. Аудио тексты должны быть расшифрованы. В этом процессе предпочтительно формировать языковой корпус из готовых электронных версий письменных текстов. Интернет может поставить неожиданные вопросы о природе языка. Он также предоставляет удобный инструмент для работы и проверки текста.

Ключевые слова: корпус, internet, вебкорпус, webCorp, NLP, BNC.

Несмотря на то, что Интернет богат лингвистической информацией на разных языках, он широко доступен и доступен бесплатно, что делает его игровой площадкой для лингвистов. Некоторые исследователи собирают данные о частоте непосредственно из коммерческих поисковых систем. Другие используют поисковую систему для поиска соответствующих страниц, извлекая текстовый материал с веб-страниц в качестве корпуса для анализа. Другая группа экспертов создает базу данных в Интернете, а затем управляет собранными данными. Таким же образом лингвистам были предложены некоторые прототипы поисковых систем Интернета. Например, веб-страница Webcorp (http://www.webcorp.org.uk/) способна генерировать словосочетания. При использовании веб-сайта в качестве корпуса у вас будут следующие возможности:

- доступ к словарю, глоссарию, тезаурусу;
- доступ к онтологиям (например, WORNET);
- анализ словосочетаний;
- Анализ фраз через поисковую систему типа Google;
- сравнительный анализ новостных сообщений в Интернете;

- создание параллельных корпусов (многие веб-страницы переведены на разные языки);
- изучение возникающих новых лексических единиц (новое употребление языка);
- изучение социальных сетей в Интернете;
- изучение специализированных корпусов (академических, деловых, новостных и других корпусов);
- изучать веб-жанры

Концепция Интернета как корпуса была использована Килгарриффом и Грефенштеттом и поставила вопрос о том, является ли Интернет корпусом. Чтобы определить, является ли Интернет корпусом или нет, было предложено сначала определить, что такое корпус. Маккенри и Уилсон заявили, что в принципе любой сборник из нескольких текстов можно назвать корпусом. Но термин корпус, используемый в контексте современной лингвистики, часто имеет более сложные значения, чем это простое определение.

Маккенри и Харди концептуализируют Интернет как корпус, во многом похожий на корпус монитора, поскольку он представляет собой большую, постоянно растущую коллекцию данных и используется для изучения языка. Помимо использования стандартных поисковых систем, таких как Google, для использования Интернета или сети в качестве базы данных, исследователи также разработали интерфейс WebCorp (Renouf, 2003), специально предназначенный для поддержки такого использования сети. WEbCorp-это веб-сайт, предназначенный для предоставления лингвистической информации. WebCorp имеет параметры поиска, специально предназначенные для лингвистических исследований (соответствия, списки слов и т. д.). Веб-сайт WebCorp предназначен для сбора лингвистической информации: существует список соответствия, показывающий контекст, в котором пользователь встретил результат поиска. Но веб-сайт WebCorp не считается полезным, поскольку он медленный и занимает некоторое время по сравнению с корпусом или поисковыми системами.

Можно сказать, что изучение сети как корпуса лингвистических исследований началось в 90-х годах, поскольку именно в те времена выросло использование интернет-текстов в качестве набора материала для корпуса. Радев и МакКаун использовали новостную сеть Интернета в качестве источника информации для своей системы генерации языков. В последние годы все исследователи используют сборники статей, доступные в Интернете,

в частности, при написании пособий, диссертаций, научных статей или получении заключений по определенным областям. Грефенштетт, Ниоке, Джонс и Гани исследовали потенциал веб-ресурсов как источника языковых корпусов для языков с небольшим количеством электронных ресурсов, а Резник исследовал потенциал параллельных двуязычных корпусов. Фудзи и Исикава написали, что использовали Интернет для создания статей в энциклопедии. Грефенштетт представил точки зрения и опыт в Интернете как источнике лексической информации, поскольку Интернет предоставляет тысячи контекстуальных примеров для многих языков, что позволяет автоматически находить лексические статьи для языков на основе эмпирических данных. Подобная область привлекает студентов и других исследователей из-за своей новизны и низких вступительных затрат. В целом приведенный выше список исследований не является исчерпывающим. Это показывает, насколько быстро он используется в качестве веб-корпуса.

Использование Интернета для корпусной лингвистики — довольно новая тенденция. Число подходов, имеющих отношение к компьютерной лингвистике, все еще невелико. Но сеть уже протестирована на задачах разного языкового уровня. Например, проводится множество исследований в области лексикографии, синтаксиса, семантики и перевода. Именно в этих четырех направлениях Волк описал конкретные аспекты подхода «Сеть как корпус». Наиболее важные задачи в лексикографии могут быть выполнены с помощью Интернета, поскольку с самого начала Интернет был источником сбора и распространения лексических ресурсов (списков слов на разных языках). Но более интересным с компьютерной точки зрения является поиск и классификация нового лексического материала из богатства текстов Интернета. Это включает в себя поиск новых слов или фраз, их классификацию и сбор дополнительной информации, такой как требования или определения для распространенных словосочетаний и подкатегорий. Может проанализировать, как выучить и классифицировать имена собственные в Интернете. Это важная работа, поскольку имена собственные представляют собой класс открытых слов, для которого постоянно изобретаются новые имена. Их система работает в три этапа. Сначала агрегатор загружает веб-страницы, полученные поисковой системой, в шаблон ключевых слов после запроса. Во-вторых, парсеры используются для поиска имен из списков и таблиц. В-третьих, модуль фильтрации очищает имена от ведущих определителей или несвязанных слов. Ряд таких задач сеть решает за

считанные минуты. В следующем направлении, синтаксисе, в Интернете есть веб-страницы, которые анализируют определенные предложения, даже более мелкие тексты. Но, по словам Волка, поисковые системы Всемирной паутины (WWW) предназначены не для лингвистических запросов, а для запросов общего характера. Сбор смысловой информации из Интернета – тоже огромная задача, аналогичная поиску лексических единиц (имен), иногда результат может оказаться неожиданным, чем ожидалось. Потому что при поиске имени его различные смысловые значения сопровождаются дополнительными синонимами и другими родственными словами. Ни для кого не секрет, что использование Интернета для услуг перевода достигло своего пика. Одной из важнейших особенностей Интернета является одновременное представление целых текстов или веб-страниц на разных языках.

С момента создания первых стандартных корпусов для английского языка-Брауна и LOB быстрые технологические изменения не только привели к распространению инструментов корпусного анализа, но и позволили добиться больших успехов в увеличении размера корпуса. Объем современных корпусов составляет не миллион, а 100 миллионов слов, и специалисты стали использовать разные материалы. Это связано с несколькими фактами:

1. Корпуса недостаточно велики для некоторых областей лингвистики, даже для новых мегаразмерных корпусов типа Британского национального корпуса (BNC). Для изучения морфологической эффективности лексических инноваций крайне необходимы материалы, отклоняющиеся от новых мегакорпораций.

2. Сами технологические разработки привели к появлению новых типов текста, ни с одним из которых корпус Брауна не сталкивался, кроме электронной почты. У них были обсуждения в чатах, текстовые сообщения, блоги или веб-журналы, и эти типы текстов являются интересными объектами изучения. Это также добавило новое измерение к типам текста для обсуждения письменных и устных слов. Потому что все они используют письменную форму, но очень близкую к образцам, которые мы видим в устной речи. Для некоторых внешних вариантов традиционные типы текста недоступны, но при обмене электронной почтой могут использоваться текстовые сообщения, блоги или интерактивные письма. Наконец, новые типы текстов во Всемирной паутине (www) представляют интерес для обсуждения социально-прагматических явлений, таких как «частный язык».

3. Создание стандартных справочных корпусов требует много времени и денег, которые в ближайшие годы быстро устареют из-за продолжающихся изменений.

4. Использование языка в Интернете само по себе может быть основным источником языковых изменений, которые постоянно меняются. Чтобы оценить влияние «веблиша» или «сетевого языка» на наш язык, нам необходимо лучше понять сам этот феномен.

Прежде всего, нам необходимо различать два способа использования Интернета в корпусных лингвистических исследованиях:

1) может использовать веб-корпус в качестве эвристического инструмента с использованием коммерческих браузеров или поисковых программ в Интернете, но также и более систематически;

2) Интернет также можно использовать в качестве ресурса для создания больших корпусов автономных мониторов (для создания веб-корпусов).

Драгомир Радев отмечает, что корпусная лингвистика, в целом обработка естественного языка (НЛП), также имеет важные услуги и возможности Интернета. Д. Радев выявил полезную разницу между «отдавать» и «получать» в механизме обработки естественного языка. НЛП рассматривает веб-технологии как веб-технологии, суммирующие машинный перевод веб-страниц или результатов веб-поиска, многоязычный поиск документов, ответы на вопросы и другие стратегии поиска не только нужного документа, но и нужной части документа, синтаксического анализа и других ключевые технологии. «Принятием или получением» считается использование вами данного веб-сайта в качестве источника информации для любых целей корпусной лингвистики или НЛП. Мы еще долго будем наполнять лингвистическую суть сети и вообще ресурсами другого назначения. Для этого нам необходимо относиться к самому сайту как к ограниченному объекту исследования. Многие технологии веб-поиска разработаны на основе языковых технологий. Кроме того, Интернет является многоязычным и инклюзивным (предполагаемые услуги), поскольку он охватывает множество языков, и в то же время его можно назвать эклектичной средой с правом выбора (возможностью выбирать лучшее). Именно эти возможности привели к оценке Интернета как корпуса.

Список использованной литературы:

1. Килгаррифф А., Грефенстетт Г. Введение в специальный выпуск Интернета как корпуса. Компьютерная лингвистика, 29 (3), 2003. стр. 333-347.

2. Терра Э., Кларк К. Оценки частоты статистических показателей сходства слов. В материалах конференции Human Language Technology и Североамериканского отделения Ассоциации компьютерной лингвистики, 2003 г., 244–251.

3. Стюарт К. Новые взгляды на корпусную лингвистику.: file:///D:/about%20corpora/Dialnet-NewPerspectivesOnCorpusLinguistics-1426958.pdf

4. МакЭнери Т., Уилсон А. Корпусная лингвистика. Издательство Эдинбургского университета, Эдинбург, 1996.

5. МакЭнери Т., Харди А. Корпусная лингвистика: Метод, теория и практика. Кембридж: Издательство Кембриджского университета, 2012.

6. http://www.webcorp.org.uk/

7. Радев Д., МакКаун К. Создание источника знаний поколений с использованием новостной ленты, доступной в Интернете. В материалах Пятой конференции по прикладной обработке естественного языка. Вашингтон, округ Колумбия, апрель 1997 г., стр. 221-228

8. Грефенштетт Г., Ниоч Дж. Оценка использования английского и неанглоязычных языков в WWW. В материалах RIAO (Recherche d'Informations Assistee par Ordinateur), Париж, 2000 г.

9. Джонс Р. и Гани Р. Автоматическое создание корпуса языков меньшинств из Интернета. 38-е заседание ACL, Материалы студенческого исследовательского семинара. Гонконг. Октябрь 2000, стр. 29-36

10. Резник П. Поиск в сети двуязычного текста. Материалы 37-го заседания ACL. Мэриленд, США, июнь 1999 г., стр. 527-534.

11. Фуджи А., Исикава Т. Использование всемирной паутины в качестве энциклопедии: извлечение описаний терминов из полуструктурированного текста. В протоколах 38-го заседания ACL, Гонконг, октябрь 2000 г., стр. 488-495

12. Грефенштетт Г. WWW как ресурс для задач MT на основе примеров. Приглашенный доклад, конференция ASLIB «Перевод и компьютер», Лондон. Октябрь 1999 года.

13. Волк М. Использование Интернета как корпуса для лингвистических исследований. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.104.6964&rep=rep1&type=pdf

14. Хундт М., Нессельхауф Н. К. Бивер. Корпусная лингвистика и Интернет. Амстердам-Нью-Йорк, 2007.