Обзор больших русскоязычных корпусов текстов

М.В. Хохлова

Санкт-Петербургский государственный университет
m.khokhlova@spbu.ru

Аннотация

В последнее время появляется все больше корпусов текстов, создаваемых автоматическими методами и отличающихся от традиционных текстовых коллекций как по объему, так и по содержанию.
В статье дается обзор корпусов русского языка RuTenTen, Генерального корпуса русского языка, корпусов проекта Aranea, а также обсуждаются вопросы, связанные с построением подобных корпусов.

Ключевые слова: корпусы текстов, Интернет-корпусы, НКРЯ, RuTenTen, RussianWeb корпус, большие данные.

Большие корпуса, объем которых превышает 100 млн слов, появились относительно недавно. И связано это с ростом технических возможностей и постепенным уходом от «ручного» составления корпусов к более автоматическому.

В настоящий момент можно говорить о корпусах двух типов (некоторые исследователи выделяют три типа [1]).

К первому типу относятся корпуса, которые создаются лингвистами по заранее описанной технологии, которую можно назвать «классической». Для подобных корпусов тексты отбираются, размечаются и далее загружаются в корпус. Самым известным и популярным корпусом первого типа для русского языка является Национальный корпус русского языка, который содержит более 600 млн слов и состоит из целого ряда подкорпусов, некоторые из них пополняются до сих пор.

Корпусы второго типа создаются автоматически на основе текстов, полученных из Интернета. Для русского языка существует ряд проектов, развиваемых в данном направлении. Наиболее ранним из них является Russian Web корпус. Данный корпус был создан С.А. Шаровым [2] на основе автоматически загруженных текстов из Интернета по технологии, описанной в [3]. Был составлен список из 500 наиболее частотных слов русского языка,

которые не являлись служебными или характерными для некоторой предметной области. На материале этого списка были последовательно заданы многочисленные запросы к поисковой системе (от 5000 до 8000), позволившие получить ссылки на сайты. В каждом запросе участвовали по четыре слова из списка. Далее для каждого запроса были отобраны 10 топ-адресов и загружены соответствующие им тексты. Затем проводилось распознавание кодировки, удаление многочисленных версий одних и тех же веб-страниц или текстов на других языках (например, в случае русскоязычного корпуса потребовалось удаление текстов на славянских языках, использующих кириллицу). Корпус был лемматизирован и морфологически размечен при помощи программы TreeTagger [4]. Общий объем корпуса составляет около 147 млн слов.

Следующие три корпуса в последние годы развивались практически параллельно.

Проект Aranea [5] получил свое название от латинского слова aranea (русск. пауки). Здесь прослеживается игра слов, т.к. технология, лежащая в основе автоматического сбора текстов (англ. crawling), использует принцип «плетения паутины», или «краулинг ползания» по Интернету. Корпуса Aranea [6] были созданы для полутора десятков разных языков, среди которых есть и русский, при помощи инструментов SpiderLing [7] и Onion [8], разработанных в Лаборатории обработки естественного языка университета им. Т.Г. Масарика (Брно, Чехия). Данные программы позволяют за сравнительно небольшое время скачать большой объем данных, извлечь из файлов текстовую информацию и удалить полные дубли и частичные повторы фрагментов. Морфологическая разметка также была осуществлена программой TreeTagger с помощью языковой модели С. Шарова [9]. В рамках проекта Aranea пользователям доступны несколько корпусов русского языка. Корпус Araneum Russicum Russicum содержит тексты, полученные с сайтов в доменах .ru и .рф. В корпус Araneum Russicum Externum вошли «внешние» русскоязычные тексты, взятые с сайтов в доменах, отличных от .ru и .рф. Корпус Araneum Russicum состоит из текстов, взятых с разных сайтов вне зависимости от доменов. Каждый из упомянутых корпусов имеет два варианта, отличающихся по своим объемам: Maius, который содержит более 850 млн слов, и Minus, содержащий 90 млн слов. Для корпуса Araneum Russicum существует еще версия Maximum, содержащая в настоящее время 10,9 млрд слов.

Семейство TenTen [10] включает в себя корпуса разных языков, объем каждого из которых превышает 1 млрд слов. Корпус русского языка ruTenTen является одним из самых больших наряду с корпусами английского, немецкого, французского и испанского языков. Особенность подхода, который используется при создании данных ресурсов, заключается в том, что скачиваются не все тексты из всех возможных доменов, соответствующих рассматриваемому языку. В качестве инструмента используется специальная поисковая программа-робот (англ. crawler), позволяющая скачивать тексты, в которых содержатся полные предложения (а не страницы с техническими данными). При этом уделяется внимание тому, чтобы новые тексты не были дублями старых, таким образом, сокращается процесс постобработки.

Для русского языка необходимо назвать также проект Генерального Интернет-корпуса русского языка, который пока содержит более 15 млрд слов,

но его создатели нацеливаются на объем в 100 млрд [11]. Планируется, что корпус будет использоваться для лингвистического анализа, поэтому большое внимание уделяется вопросам, связанных с метаразметкой текстов. Морфологическая разметка текстов была проведена при помощи программы TnT-Russian [12]. Пока в корпусе преимущественно представлены тексты social media — блогов, социальных сетей, форумов. На сегодняшний день это единственный большой корпус, который задумывался изначально только для русского языка, а не являлся частью многоязычного проекта.

Упомянутые выше технологии являются привлекательными, так как позволяют создавать корпуса для разных языков, не требуя предварительного затратного по времени и силам сбора текстов (хотя это достоинство можно поставить под сомнение, вспомнив о таком неотъемлемом свойстве корпусов и выборок как сбалансированность, или репрезентативность).

Направление, связанное с построением и использованием больших корпусов текстов является весьма перспективным в корпусной лингвистике. В последнее время подобные корпуса становятся материалом для разных исследований, в том числе по сравнению данных коллекций текстов между собой.

Статья подготовлена в рамках работы по гранту Президента Российской Федерации для государственной поддержки молодых российских ученых № МК-5274.2016.6 «Исследование статистических закономерностей сочетаемости лексических единиц».

Литература

[1] Беликов В.И., Селегей В.П., Шаров С.А. Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ) 2012 // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). М., 2012. Т. 1. С. 37–49.

[2] Sharoff S. Creating General-Purpose Corpora Using Automated Search Engine Queries // WaCky! Working papers on the Web as Corpus. Bologna, 2006.

[3] Baroni M., Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web // Proceedings of LREC 2004. Lisbon: ELDA, 2004. P. 1313– 316.

[4] Schmid H. Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. UK, Manchester, 1994. P. 44–49.

[5] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora // Proceedings of Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. LNCS 8655. Springer International Publishing Switzerland, 2014. P. 257–264.

[6] Aranea. URL: http://sketch.juls.savba.sk/aranea_about/ (дата обращения: 25.04.2016).

[7] Suchomel V., Pomikalek J. Efficient Web Crawling for Large Text Corpora // Proceedings of the Seventh Web as Corpus Workshop (WAC7). Lyon, 2012. P. 39–43.

[8] Pomikalek J. Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis, Masaryk University, Faculty of Informatics, 2011.
[9] Russian Statistical Taggers and Parsers. URL: http://corpus.leeds.ac.uk/mocky/ (дата обращения: 25.04.2016).
[10] Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. The TenTen Corpus Family // Proceedings of the 7th International Corpus Linguistics Conference. Lancaster, 2013. P. 125–127.
[11] Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S. Big and Diverse is Beautiful: A Large Corpus of Russian to Study Linguistic Variation // По материалам Web as Corpus Workshop (WAC–8). URL: http://www.webcorpora.ru/wp-content/uploads/2015/10/wac8-proceedings.pdf (дата обращения: 25.04.2016).
[12] Sharoff S., Nivre J. The Proper Place of Men and Machines in Language Technology: Processing Russian without any Linguistic Knowledge // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25–29 мая 2011 г.). Вып. 10 (17). М.: РГГУ, 2011. С. 657–670.

A survey of Large Russian Corpora

M. Khokhlova
St. Petersburg State University

Our paper deals with large corpora rapidly appearing within the last decade. These corpora are created automatically and differ from traditionally made text collections both in their volume and content. The paper gives a survey of large Russian corpora, among them RuTenTen, General Internet-Corpus of Russian, project Aranea and also discusses the issues of building such corpora.

Keywords: text corpora, web corpora, Russian National corpus, RuTenTen, Russian Web corpus, big data.