Секция «Востоковедение, африканистика»

Типология ошибок тэгирования корпуса монгольского языка

Дамбиев Зандан Баирович
Студент
Бурятский государственный университет, Национально-Гуманитарный Институт, Улан-Удэ, Россия
E-mail: haeywo@gmail.com

Лингвистическим корпусом принято называть совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Целесообразность создания корпуса объясняется возможностью многократного использования единожды созданного корпуса для решения различных лингвистических задач, таких, как например, реализация графематического и лексико-грамматического анализа текста, представлением лингвистических данных в реальном контексте и при достаточно большом объеме корпуса - высокой достоверностью данных. Корпус состоит из конечного числа текстов, но он призван адекватно отражать лексико-грамматические феномены, типичные для всего объема текстов в соответствующем языке (или подъязыке). Для представительности важен как размер, так и структура корпуса. Представительный размер зависит от задачи, поскольку он определяется тем, как много примеров может быть найдено для исследуемых феноменов[2, 2]. В качестве эксперимента нами был создан учебный корпус монгольского языка, состоящий из 10 000 словоупотреблений. При обработке корпуса нами была использована программа AntConc. В процессе обработки результатов тегирования был выявлен ряд типичных ошибок. Первая, наибольшая группа ошибок была связана, прежде всего, со спецификой строя монгольского языка, одной из особенностей которого является синкретизм основ. Так, единица *балга* в одном случае была распознана как глагольная основа от *балгах* «глотать», а в другом случае эта же единица *балга* была репрезентована как имя существительное «глоток». Следующая группа некорректностей была вызвана относительно узким диапазоном запросов к поисковой машине. Например, определенные сложности возникли с соотнесением единиц к тем или иным частям речи: единица *х1257;нг1257;н шуурхай* «быстро, быстрый» был отнесен только к именам прилагательным, несмотря на то, что в виду все той синкретичности, является и наречием. Безусловно, представленный учебный экспериментальный корпус монгольского языка не претендует на представительность, однако он вполне пригоден для решения ограниченного круга лингвистических задач. Кроме того, после обнародования результатов анализа ошибок первого корпуса монгольского языка, мы пришли к выводу, что выявленные нами некорректности являются типичными. В 2008 году группой монгольских педагогов - лингвистов из Монгольского Государственного Университета был проведен анализ ошибок корпуса монгольского языка, состоящего из пяти миллионов словоупотреблений. Был создан набор тегов частей речи монгольского языка. Согласно проведенному анализу было выявлено 100 тысяч тегированных слов, содержащих ошибки при описании. При этом было выявлено три основные группы ошибок: 1) первые 15 тысяч слов, ошибки при обработке которых, были вызваны слишком общим характером тегов (например, в определенном контексте *б1199;х* «целый»был тегирован как прилагательное, хотя не было уточнено, что оно непроизводное, ***нэг нь***

нийтийн т1257;л1257;1257; «один за всех» был тегирован как существительное, хотя является числительным); 2) следующие 62 тысяч слов, в которых встретилась проблема ошибочного тегирования частей речи, эти слова были вынужденно тегированы вручную (например, наречие *амжилттай* было ошибочно тегировано как прилагательное); 3) последние 27 тысяч слов, здесь сыграл свою роль человеческий фактор, неверный анализ либо ошибки тегирования (например, относительное прилагательное *голын* было тегировано как имя нарицательное в родительном падеже). Тегированный корпус был проанализирован с помощью конкорданс программ AntConc и MTagger. Для анализа и выявления ошибок в тегированных текстах, был создан конкорданс текста с тегами частей речи и слов, а также созданы различные виды таблиц и списков, показывающих статистическую информацию о тегированном корпусе. Монгольские лингвисты считают, что анализ такого вида информации является основным методом анализа корпуса. Анализ и обслуживание корпуса, равно как его создание, требует много работы. Соответственно, необходимо определить принципы анализа корпуса и его поддержания после тегирования корпуса. Это занимает больше времени и требует больше опыта от лингвистов, так как это первый корпус монгольского языка. Таким образом, разрабатываются рекомендации, основанные на опыте и анализе тегирования 100 тыс. слов, о том, как анализировать и поддерживать тегированный корпус. Большинство ошибок, допущенных во время тегирования корпуса, допущены во время разработки набора тегов частей речи. Монгольский язык является одним из агглютинативных языков. В синтаксисе монгольского языка, словоизменительные функции имеют как аффиксы, так и изолированные слова (разделенные пробелами орфографически, но несущие словоизменительные функции, так же, как в случае с аффиксами и показателями множественного числа). Кроме того, существует много неясностей между морфологической структурой и синтаксисом позиции для одной формы. Например, *-тай* является признаком аффикса, как флективным так и деривационным, так что с ним могут возникнуть вопрос, каким именно аффиксом он является в данном случае. Это может быть прилагательное либо модальное слово, иметь морфологические и синтаксические особенности. Допустим, *х1199;чтэй* может рассматриваться как существительное «сила» в форме совместного падежа с аффиксом *-тэй*, либо как прилагательное «сильный», «мощный». Таким образом, в настоящее время описанные выше проблемы затрудняют создание полного набора тегов частей речи монгольского языка.

Литература

1. Kilgarriff A. Googleology is bad science. Computational Linguistics, 33(1), 2007
2. Бочаров В.В., Грановский Д.В. Программное обеспечение для коллективной работы над морфологической разметкой корпуса (рус.) // Труды международной конференции «Корпусная лингвистика – 2011». — Санкт-Петербург: СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011
3. Корпусная лингвистика: http://corpora.iling.spb.ru/
4. PANLocalization Project, Error analysis of mongolian corpus, // National University of Mongolia, Ulaanbaatar, Mongolia, 2008