Гатиатуллин Айрат Рафизович
Кубединова Ленара Шакировна
Прокопьев Николай Аркадиевич
Абдураманов Ибраим Арсенович
Институт прикладной семиотики,
Академия наук РТ
УДК 004.822

Инструментарий портала «Тюркская морфема» для создания электронных корпусов тюркских языков в рамках единого концептуального пространства

электронный корпус, концептуальное пространство, лингвистический ресурс, тюркология

1. Введение

Актуальность и важность создания электронных корпусов естественных языков как инструмента их сохранения и развития, а также как ресурсной базы для разработчиков NLP-технологий и исследователей языков, у специалистов не вызывает сомнения. В настоящее время благодаря возрастающей поддержке со стороны государственных органов, различных международных фондов и большому вниманию к проблемам исчезающих и малоресурсных языков со стороны ЮНЕСКО, наблюдается быстрый рост количества и объема электронных корпусов для множества языков, в том числе и для тюркских языков [UNESCO 2019]. Для многих из языков тюркского семейства существует уже по несколько электронных корпусов.

Вместе с тем, существует ряд проблем в процессе создания электронных корпусов, решению которых направлены представленные в статье исследования. В частности, адресована проблема создания новых корпусов для некоторых тюркских языков, так как не у всех желающих создать подобные ресурсы, имеется такая возможность. Разработчики сталкиваются с техническими проблемами размещения и поддержки функционирования таких крупных, требовательных к техническому обеспечению ресурсов как корпусы, а также со сложностями программной реализации собственного инструментария для разметки электронных корпусов с учетом особенностей языка.

2. Подходы к решению проблемы

Создание электронных лингвистических корпусов выдвигает перед разработчиками широкий спектр проблем и задач, успешное решение которых требует соединения результатов лингвистических исследований и современных компьютерных методов анализа языковых данных.

Анализ сложившейся ситуации [Kubedinova, Adali 2019] показывает, что в тюркской корпусной лингвистике, несмотря на генетическую и структурно-типологическую общность тюркских языков, пока не сформировалось общих принципов и подходов к лингвистической разметке текстов. По заключению участников конференции TurkLang [Psyanchin 2020], одной из важнейших задач тюркского языкознания является выработка единого стандарта представления лингвистической информации, который позволяет интегрировать существующие и создающиеся корпуса тюркских языков в единое информационное пространство. В качестве дополнения к организационным мерам необходимо иметь единые ресурсные базы и инструментарий. Такой ресурсной базой может стать портал «Тюркская морфема» [Gatiatullin 2020] (modmorph.turklang.net).

3. Инструментарий для создания корпусов

Инструментарий для создания новых корпусов в рамках портала «Тюркская морфема» обеспечивает структуру базы данных, которая пригодна для большинства корпусов. Сущности этой базы данных связаны с языковыми элементами Модели Тюркской Морфемы на основных уровнях: грамматическом, синтаксическом, семантическом.

База данных нового корпуса содержит пять основных сущностей для хранения корпусных данных на различных уровнях детализации:

1. «Документ» – хранит данные об отдельном текстовом документе корпуса, Эта сущность связана с сущностью «Язык» из базы данных портала.

2. «Предложение» – хранит данные об отдельных предложениях из текста «Документа».

3. «Клауза» – хранит данные о клаузах предложений из текста документа. Каждая запись сущности «Клауза» соответствует некоторой записи сущности «Ситуационный фрейм» из базы данных портала.

4. «Синтаксема» – хранит данные о синтаксемах клауз, в рамках портала записи данной сущности связываются с записями сущности «Ситуационная роль».

5. «Словоформа» – хранит данные об отдельных словоформах синтаксем. Представляет собой отдельное слово, разобранное при помощи морфологического анализатора. Результат разбора связывает записи данной сущности с записями сущностей «Корневая морфема», «Аффиксальная морфема», «Аналитическая морфема». Корневые морфемы, в свою очередь, связаны с сущностью «Концепт». Для каждого нового корпуса создается отдельное типовое веб-приложение, называемое модуль корпуса.

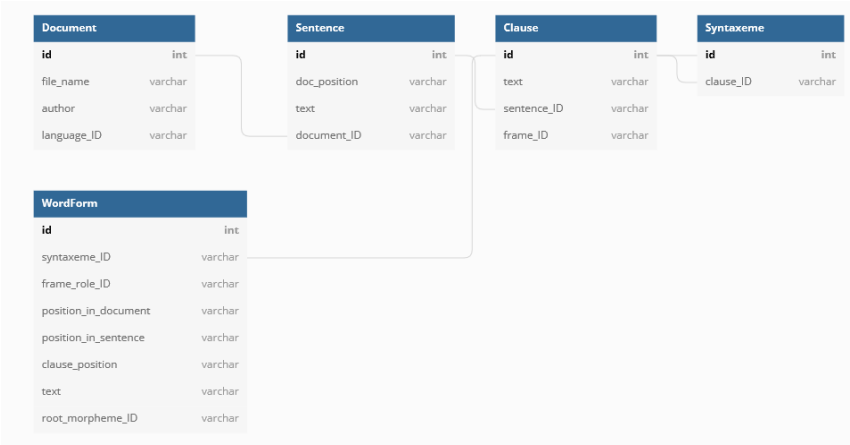
Каждый модуль корпуса связывается с базой данных портала через предоставляемый единый программный интерфейс (API). Кроме того, модули корпуса реализуют собственный единый API, который обеспечивает доступ к данным при помощи поисковых запросов и позволяет производить выгрузку данных.

4. Реализация модуля корпуса

Каждый модуль корпуса реализован как типовое веб-приложение на кроссплатформенном интерпретируемом языке Python с использованием фреймворка Django. Базовая кодовая и модульная структура такого шаблона может быть загружена на сервер пакетного менеджера pip в виде загружаемой библиотеки, таким образом упрощая возможную дальнейшую разработку. Далее рассмотрен пример создания нового корпуса на основе данного инструментария.

В качестве СУБД выбрана PostgreSQL. Модуль корпуса создает базу данных с таблицами, представленными на схеме на рисунке 1.

Рисунок 1. Физическая ER-диаграмма модуля корпуса



Пользовательский интерфейс модуля дает доступ для экспертов-лингвистов к заполнению данных корпуса с использованием автоматизированных средств. Каждая сущность базы данных модуля корпуса доступна для редактирования в интерфейсе.

После загрузки документа инициализируется алгоритм разбиения документа на предложения, которые записываются в базу данных модуля корпуса. Алгоритм основан на использовании библиотеки для обработки естественного языка NLTK. Сначала из документа извлекается текст с использованием библиотеки textract, позволяющей обрабатывать большинство текстовых и офисных форматов документов. Для разбиения на предложения используется функция nltk.sent_tokenize, дающая на выходе набор предложений, сохраняющихся далее в базу данных с вычисленной позицией в документе.

5. Заключение

В статье представлен инструментарий для создания новых корпусов корпусов тюркских языков в рамках портала «Тюркская Морфема», рассмотрена реализация отдельного модуля корпуса. Использование данного инструментария позволяет поддерживать единую иерархическую структуру корпусной базы данных, связанную с общеязыковыми и языкозависимыми ресурсами портала и обеспечивает общий программный интерфейс доступа к данным для разработки и подключения программного обеспечения, такого как корпус-менеджер, инструменты для обработки естественного языка, графические интерфейсы для выгрузки и представления сводных данных по нескольким языкам и геолингвистические системы.

Литература

Gatiatullin A. About Turkic Morpheme portal, in Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) / A. Gatiatullin, D. Suleymanov, N. Prokopyev, B. Khakimov // CEUR-WS, 2020. – Pp. 226–243.

Kubedinova L., Adali E. The Crimean Tatar Electronic Corpus vs the Electronic Corpus of the Turkish Language (Grammatical Tagging of Noun, Verb) / L. Kubedinova, E. Adali // in Proceedings of 2019 4th International Conference on Computer Science and Engineering (UBMK), 2019. – Pp. 783–788.

Psyanchin A.V. (Ed.) VIII International Conference on Computer Processing of Turkic Languages “TurkLang 2020”, Proceedings, IHLL UFRC RAS. – Ufa, Russia. – 2020.

UNESCO Headquarters International Conference Language Technologies for All (LT4All): Enabling Linguistic Diversity and Multilingualism Worldwide in the framework of the 2019 International Year of Indigenous Languages 4–6 December 2019. – Paris, France.