Проектирование базы данных для разметчиков параллельного корпуса технических текстов

Рунов Константин Алексеевич runovka@student.bmstu.ru

МГТУ им. Н. Э. Баумана, Москва, 105005, Россия

В статье рассмотрены актуальные проблемы разметчиков параллельных корпусов технических текстов, рассмотрена функциональность информационной системы для решения некоторых из них, представлена диаграмма сущностей базы данных для описанной информационной системы, и приведен сценарий использования системы

Ключевые слова: *корпусная лингвистика, параллельные корпуса текстов, терминологическая разметка, база данных*

Design of database for parallel corpora of technical texts annotators

Runov Konstantin Alexeevich runovka@student.bmstu.ru

BMSTU, 105005, Russia

The article considers actual problems of parallel corpora of technical texts annotators, considers the functionality of the information system for solving some of them, presents the diagram of database entities for the described information system, and gives a scenario of the system usage

Keywords: corpus linguistics, parallel corpora, terminological markup, database

Введение

На сегодняшний день корпусная лингвистика является неотъемлемой частью лингвистики, науки о языке [1]. Корпуса текстов находят применение в различных областях — в машинном переводе, в разработке словарей, в лингвистических исследованиях. Для того, чтобы из корпуса текстов можно было извлекать пользу, тексты в нем должны быть размечены. Существуют различные алгоритмы, позволяющие автоматически производить разметку текстов, но для проверки ее корректности все равно, как правило, требуется вмешательство человека.

На данный момент не существует открытых параллельных корпусов технических текстов, а также инструментов для автоматической разметки русскоязычных текстов [2, 3]. Это делает актуальным разработку информационной системы для участников разметки параллельного корпуса технических текстов.

Целью данной статьи является создание диаграммы сущностей базы данных для информационной системы, предназначенной для участников разметки параллельного корпуса технических текстов.

Для достижения поставленной цели, будут

— рассмотрены актуальные проблемы, с которыми сталкиваются участники разметки параллельных корпусов технических текстов,

— описана функциональность информационной системы, позволяющей организовать удобную работу множества участников разметки параллельного корпуса технических текстов,

— выделены сущности базы данных для описанной информационной системы.

Проблемы разметки параллельных корпусов технических текстов

Актуальные проблемы разметки параллельных корпусов технических текстов связаны с автоматическим определением границ терминов. Среди них выделяют [3]:

— неправильное определение границ терминов-словосочетаний, состоящих из двух и более слов и составных терминов;

— распознавание составных терминов и терминов-словосочетаний, состоящих из двух и более слов;

— определение лексической единицы как термина в зависимости от контекста и тематики текста, в котором данная лексическая единица употребляется;

— объемные списки терминов-кандидатов, которые необходимо проверять вручную.

Точность определения границ термина при автоматической разметке является одной из основных лингвистических задач, но на сегодняшний день отсутствуют информационные системы для автоматической разметки русскоязычных текстов [3].

Существующие параллельные корпуса текстов и информационные системы для разметки текстов

Существует множество параллельных корпусов (My-Memory, Opus, Linguee, Glosbe, Reverso, TAUS Data Cloud и др.) [4], сервисов для автоматического выравнивания текстов (Hunalign, Euclid, Abbyy Aligner, Trados, Winalign, Wordfast tools, Giza++ и др.) [1], инструментов для автоматического извлечения терминов (TerMine, TermExtraction, Terminology Extraction) [3]. Однако не существует открытых параллельных корпусов технических текстов, а также инструментов для проведения автоматической разметки русскоязычных текстов.

Функциональность информационной системы для участников разметки параллельного корпуса технических текстов

Для организации удобной работы множества участников разметки параллельного корпуса технических текстов, информационная система должна обладать рядом функций, которые можно разделить на три категории — функции пользователя, функции модератора и функции администратора.

Администратор системы имеет полный доступ к данным: может добавлять, удалять тексты, добавлять, удалять, изменять разметки, назначать новых модераторов.

Модератор может создавать, изменять и удалять задания на разметку, а также производить проверку разметки, после чего либо утверждать, либо отклонять ее.

Пользователь может выполнять задания на разметку, просматривать выполненные разметки, производить поиск по корпусу.

Каждый пользователь имеет свой рейтинг доверия, который растет с увеличением количества его утвержденных разметок.

Сущности базы данных для описанной информационной системы и сценарий использования

На рисунке 1 предложен вариант диаграммы сущностей базы данных для описанной информационной системы.

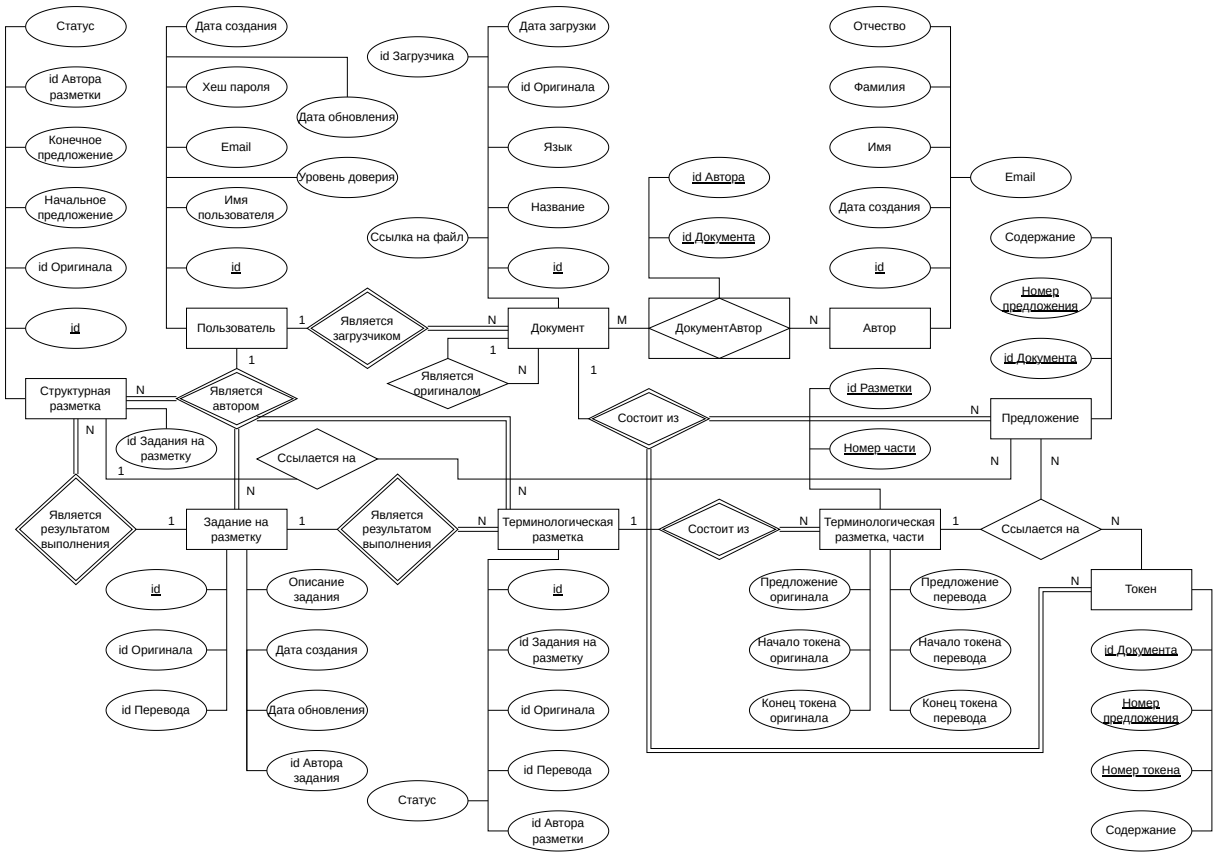

Рисунок 1 – ER-диаграмма в нотации Чена

Сценарий использования системы может быть следующий:

1) Администратор загружает документ в систему, перед этим заполнив метаинформацию о нем — язык, название текста, информацию об авторах.

2) Из документа извлекаются предложения и добавляются в таблицу «Предложение» базы данных со ссылкой на документ, которому они принадлежат.

3) Производится токенизация загруженного текста, токены добавляются в таблицу «Токен» базы данных, со ссылкой на документ и предложение, которым токен принадлежит.

4) Модератор создает задание на разметку, указав документы (оригинал и перевод), которые требуется разметить, а также добавив словестное описание задания.

5) Пользователь видит новое задание и приступает к разметке — сопоставляет части текста одного документа с соответствующими им частями текста другого документа.

6) По завершении разметки, пользователь отправляет ее на проверку модераторам.

7) Модераторы проверяют разметку пользователя и либо утверждают, либо отклоняют ее. В случае утверждения разметки, у пользователя увеличивается рейтинг доверия.

Заключение

В данной статье были рассмотрены актуальные проблемы, с которыми сталкиваются участники разметки параллельных корпусов технических текстов, описана функциональность информационной системы, позволяющей организовать удобную работу множества участников разметки, выделены сущности базы данных для описанной информационной системы и приведен сценарий ее использования.

Список литературы

[1] Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. — СПб.: Изд-во С.-Петерб. ун-та, 2020. — 234 с.

[2] Бутенко Ю.И., Строганов Ю.В., Бабаджанян Р.В. Исследовательский прототип параллельного корпуса научно-технических текстов // Актуальные проблемы лингвистики и лингводидактики в неязыковом вузе: 4-я Международная научно-практическая конференция. 2020. Т. 1. С. 205–209.

[3] Бутенко Ю.И., Сулейманова Э.В. Терминологическая разметка научно-технических текстов в специальном корпусе // Проблемы лингвистики и лингводидактики в неязыковом вузе: 5-я Международная научно-практическая конференция. 2022. Т. 1. С. 329–337.

[4] Бутенко Ю.И., Киселёва А.Д. Анализ современных корпусов параллельных текстов // Актуальные проблемы лингвистики и линг-

водидактики в неязыковом вузе: 4-я Международная научно-практическая конференция. 2020. Т. 1. С. 238–242.