Блок “**REALEC-сайт**”

Актуальность

Это направление работы посвящено совершенствованию сайта корпуса REALEC и расширению функциональных возможностей корпуса.
Корпус Russian Error-Annotated Learner English Corpus -- динамично развивающийся проект, который используется в преподавании английского языка в Школе лингвистики Высшей школы экономики. Пользовательская аудитория нашего корпуса постоянно растет, и очень важно, чтобы сайт корпуса развивался вместе с пользователями и соответствовал нуждам преподавателей и студентов.
Существовавший ранее интерфейс корпуса был предназначен исключительно для разметки и просматривания работ студентов. Все исследования по материалам размеченных текстов требовали предварительного отбора материала и его анализа вне интерфейса корпуса. К тому же рамках научно-учебной группы нами были разработаны лексические тренажеры для изучающих английский язык. Создание веб-приложения для онлайн-тренажеров -- трудоемкий процесс, и ранее мы создавали тренажёры для студентов на онлайн-сервисах. Теперь было необходимо перенести тренажёры из онлайн-сервис в нашу систему.
Для того чтобы не создавать специальное веб-приложение для тренажеров с нуля, мы решили изучить готовые программные решения, в которых предусмотрена возможность загрузки тренажеров. Речь о результатах пойдет в следующем разделе.

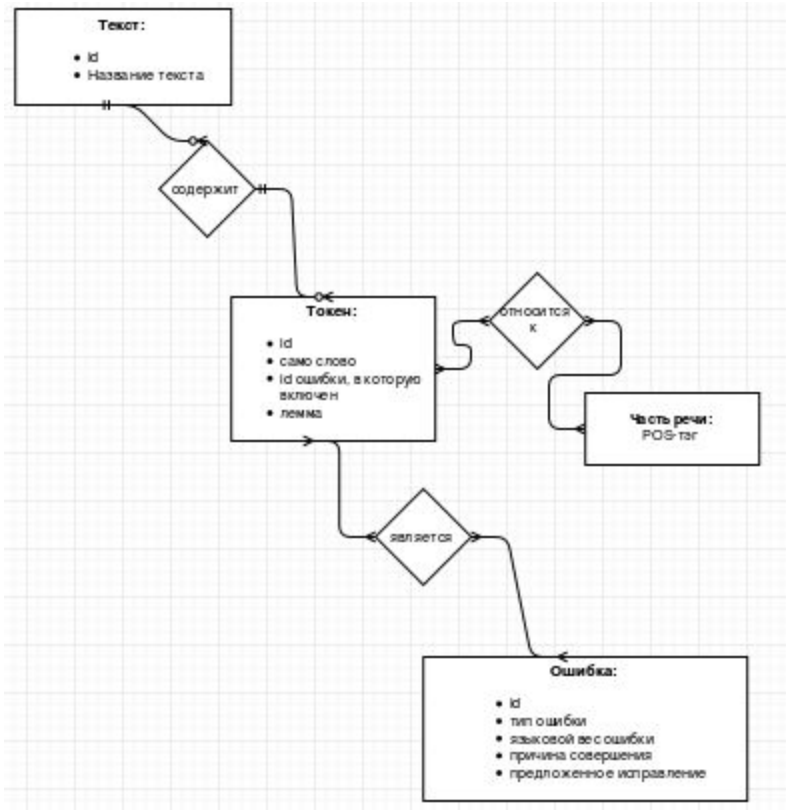
Было сделано

Мы остановили наш выбор на системе управления обучением Moodle. Сайт, созданный с использованием Moodle, находится по адресу http://web-corpora.net/realec/. Помимо загрузки тренажеров, Moodle позволяет также настроить дифференциацию групп пользователей. На платформе предусмотрены следующие группы пользователей:

- Студенты;
- Преподаватели;
- Внешние пользователи без логина.

Студенты могут проходить упражнения и просматривать результаты пройденных тестов. Преподаватели могут видеть оценки студентов, выкладывать новые материалы, а также редактировать тесты и структуру курса в системе Moodle. Внешние пользователи могут просматривать материал, но не проходить его.
Помимо создания дополнительного ресурса на платформе Moodle, мы также усовершенствовали поисковую систему для корпуса. Ранее поиск по корпусу мог быть только текстовым: пользователь вводит фразу или название тэга, и программа ищет соответствия заданной цепочке символов во всех текстовых файлах и файлах с разметкой. При таком подходе информация о разметке ошибок в корпусе никак не была структурирована для поиска, и поиск занимал много времени и машинных ресурсов, в то время как его возможности были не очень велики. Мы сделали первые шаги в

совершенствовании поиска в корпусе, хотя этот процесс ещё не завершен. Мы переработали файлы с разметкой и поместили их в реляционную базу данных. Структура базы данных такова:



Единицей в базе данных является текст, поскольку корпус состоит из отдельных текстов. Тексты обладают такими свойством как название. Каждый текст состоит из токенов. Важно, что токены располагаются не в случайном порядке в таблице, а в порядке следования в тексте. Тогда, например, легко можно будет извлекать контекст нужной длины, взяв какое-то количество предшествующих и следующих в таблице токенов. Каждый токен обладает двумя свойствами: во-первых, он принадлежит какому-то из частеречным тэгом (точнее, в некоторых случаях кроме части речи есть ещё и какая-то морфологическая информация, но это английский язык, так что морфология бедная). Второе свойство -- принадлежность к области ошибки. Соответствующую ошибку можно указывать по её id, а если токен ни в какую ошибку не включается, то тогда, наверно, можно ставить 0. Каждый токен может быть частью как одной, так и нескольких ошибок. Последняя сущность в базе -- ошибка. Ошибка состоит из токенов и обладает следующими свойствами:

- тип ошибки
- вес ошибки с точки зрения языка
- вес ошибки с точки зрения понимания текста
- причина ошибки
- является этот фрагмент удалением или вставкой в оригинальный текст
- предложенное исправление

Тогда, например, если мы ищем ошибки конкретных типов, мы сначала отбираем все подходящие сущности из таблицы ошибок, а потом смотрим на дополнительные ограничения (частеречные тэги и контекст), для этого идем в таблицу с токенами и отбираем те ошибки, которые удовлетворяют ограничениям.
Прототип поиска на основе реляционной базы данных можно найти по ссылке http://realec.org/search/ В настоящий момент мы также тестируем возможность графового поиска: все тексты представляются в виде узлов графа, связанных каким-либо отношением (например, содержащие ошибки одинакового типа), и поиск производится по этому графу
Помимо усовершенствования поисковой системы в корпусе, мы также начали разрабатывать возможность хранения метаданных о содержащихся в корпусе текстах. До настоящего момента в корпусе содержались только сами тексты и данные о размеченных в них ошибках. Часть текстов была также размечена морфологически. Однако отсутствие метаразметки исключает возможность проведения социологических исследований, например, изучения разницы в ошибках, допускаемых юношами и девушками, или ошибок, специфичных для какого-либо направления обучения (например, ошибки студентов-историков или ошибки студентов-экономистов). Таким образом, добавление метаразметки в корпус REALEC является актуальной задачей. Мы разработали следующий формат метаразметки:

- sex - пол учащегося (f, m)
- mark - оценка за работу (от 0 до 10)
- study_year - курс студента (от 1 до 4)
- date - дата написания работы, год в самом простом случае
- department - специальность
- IELTS - написана ли работа при подготовке к IELTS или его сдаче (True/False)
- work_type - экзамен/контрольная/домашняя/работа в классе (exam, test, hw, class_work)
- text_type - несколько макротипов текстов (opinion essay, graph description, review, etc.)

Эти данные записываются для каждого текстового файла в корпусе и хранятся в формате json. В будущем мы планируем также добавить в поисковый механизм возможность искать информацию в метаданных.
К сожалению, данные о содержавшихся ранее в корпусе эссе безвозвратно утеряны, однако эссе, написанные в рамках тестирования IELTS, проводимого ВШЭ, снабжены метаразметкой.

Обобщение

Итогом усовершенствования сайта корпуса REALEC в рамках деятельности НУГ “REALEC для реально необходимых слов” стало следующее:

- Создание платформы для онлайн-тренажеров на основе системы Moodle. Платформа дифференцирует различные группы пользователей.

- Разработка новой поисковой системы для корпуса REALEC, замена текстового поиска на поиск по базе данных и графовый поиск
- Разработка формата для хранения метаданных о материалах корпуса.

Новые функциональные возможности, созданные для нашего корпусного ресурса, предоставляют пользователям корпуса возможность проводить новые типы исследований и совершенствовать знание английского языка.