CAT and kittens

Русский академический корпус

Руководители:
Копотев Михаил
Зевахина Наталья
Kisselev Olesya
Толдова Светлана

Участники:
Баранчикова Анастасия
Дмитриева Анна
Климов Александр
Фёдорова Мария

“Всё страньше и страньше! Всё чудесатее и чудесатее! Всё любопытственнее и любопытственнее! Всё страннее и страннее! Всё чудесится и чудесится!”

Описание проекта

Онлайн-сервис, который сравнивает студенческий текст (“kitten”) с эталонным представительным корпусом (CAT) и помечает «подозрительные» речевые единицы (токены, коллокации, конструкции), которые **резко** отличают студенческий текст от эталонных.

“— Наверно, я бы лучше поняла, — сказала Алиса чрезвычайно учтиво, — если бы это было написано на бумажке, а так я как-то не уследила за вами”.

Цель

Создание сервиса CAT and kittens.

Главные задачи сервиса: а) помощь иностранным студентам в освоении русского языка; б) помощь русским студентам в написании академических текстов; с) создание основы для написания учебников и справочников по академическому письму.

“Лучший способ объяснить — это самому сделать!”

Задача первая

Создание корпуса CAT:

1. выработка критериев отбора текстов
2. выработка структуры метаданных (с учетом уже имеющихся на сайтах cyberleninka.ru и elibrary.ru)
3. получение текстов (crawling) и метаданных (page parsing) на основе выработанных критериев
4. собственно создание корпуса (включая лемматизацию, аннотирование и т.д.)

“Лучший способ объяснить — это самому сделать!”

Задача вторая

Создание онлайн сервиса:

• выработка параметров для сравнения kitten относительно CAT (например, длина предложения; частоты знаменательных токенов / n-грам; близость текста к эталонным по моделям doc2vec / word2vec; вхождение (ключевых) токенов в семантический класс соответствующего тематического домена и др.)

• Создание системы сравнения по выработанным параметрам: а) создание сервиса, который «на лету» создает для загруженного kitten: семант. вектора, n-граммы, лемматизацию и т.д. и б) создание алгоритмов сравнения этих параметров с данными из CAT «эталонными».

“— А что это за звуки, вон там? – спросила Алиса, кивнув на весьма укромные заросли какой-то симпатичной растительности на краю сада.
— А это чудеса, – равнодушно пояснил Чеширский Кот.
— И.. И что же они там делают? – поинтересовалась девочка, неминуемо краснея.
— Как и положено, – Кот зевнул. – Случаются...”

Что сделано?

1. Собрана команда! Проведен первый созвон.
2. Есть выгруженный из Киберленинки корпус академических текстов, который, однако, нужно проверить на соответствие критериям корпуса, который мы собираем (актуальность материалов, темы, наличие\отсуствие метаразметки).
3. Есть список доменов (социология, история и т. д.) к которым должны относиться тексты.

“ Если бы это было так, это бы ещё ничего. Если бы, конечно, оно так и было. Но так как это не так, так оно и не этак. Такова логика вещей”.

Что не сделано?

“Ничего не сделано” (с) М. Копотев

“Если даже есть талант,
Чтобы не нарушить, не расстроить.
Чтобы не разрушить, а построить.
Чтобы увеличиться, удвоить и утроить -
Начерти на карте план!”

Какой план?

1. Выработать требования к текстам и написать ТЗ;
2. Посмотреть, что сделано для других языков;
3. Собрать и проаннотировать корпус к Новому году;
4. К Новому году сделать первый вариант будущей оболочки сервиса;
5. После Нового года выработать критерии сравнения студенческих текстов с эталонными;
6. До конца учебного года реализовать систему сравнения (back-end);
7. В следующем учебном году выложить окончательный доступный для всех вариант сервиса, например, на webcorpora (front-end).

“План, что и говорить, был превосходный; простой и ясный, лучше не придумать. Недостаток у него был только один: было совершенно неизвестно, как привести его в исполнение”.

Что непонятно?

- Как добиться репрезентативности корпуса?
- Как оценивать качество текстов?
- Как будет проходить подготовка текстов к автоматической разметке?
- Каковы будут критерии автоматической разметки?