Встреча 03.01.2018

Обсудили регулярность встреч. Встречаемся по средам в 22:30.

Рассказали, сколько и какой объём текстов мы уже собрали для корпусов (пока есть очевидное смещение в область московских и петербургских университетов). Обсудили сбалансированность корпусов (втч необходимость включения учебников в корпус).

Обсудили разметку: главное - разметка морфологии, но разметка синтаксиса нужна тоже. Используем TNT либо *[самый оптимальный вариант]* http://web-corpora.net/wsgi3/ru-syntax/ (нужно написать разработчикам (https://ufal.mff.cuni.cz/kira-droganova) на случай, если парсер был улучшен).

Обсудили параметры проверки (возможно, движемся в направлении, предлагаемом Sketch Engine):

1) Частотность слов: появления терминов (по частям речи), которые не встречаются в доменах (или в общенаучном корпусе).
2) Определение регулярно используемых устойчивых биграмм или триграмм в тексте, которые не встречаются в доменах - по стандартным коллокационным мерам (оптимально - частоты и T-score/MI и их сочетание + частеречные фильтры; см. статью http://www.helsinki.fi/~kopotev/Pivovarova_Kormacheva_Kopotev-2017.pdf; skip-gram??? поиск не только стоящих рядом, но и сочетаний: например, фиксировать глаголы и существительное в определённом падеже; возможно, использовать word2vec: “после такого глагола мы ожидаем такие-то элементы”);
3) ??? Лексическое разнообразие (богатство) в тексте: пропорция разнообразных слов (Type-Token Ratio). С этим есть проблема, т.к. нет конкретных указаний. + соотношение частей речи (“В научной речи превалируют генитивные цепочки”, “В научной речи употребляется меньше глаголов”). Это может быть полезно для преподавателей, которые могли бы оценивать своих студентов и их освоение лексики (?). Результат можно выдавать в виде таблички (“ваш текст по TTR близок к эталону на …”; “Ваш текст по Word2Vec…”). Можно оценить частотные глаголы и предлагать синонимы по Word2Vec (например, когда злоупотребляют глаголом “быть”). Для этого нужно учитывать контексты и коллокации.

4) (?) Просмотр синтаксических конструкций: пассивных, сравнительных, модели управления (глагол + падеж, можно использовать LL, или Freq.ratio), согласование (прилагат+сущ.) и т.д.

Смотреть на ошибки в этих конструкциях. Сравнивать стоит не с отдельными доменами, а со всем корпусом в целом.

5) По КРУТ можно посмотреть, какие синтаксические ошибки делаются регулярно: можно посмотреть, как часто эти синт. структуры встречаются в наших доменах. Из КРУТ извлекаются кластеры ошибок, хотя по тэгам может быть много мусора. Надо найти 3-4 самых частотных конструкции с ошибками или 4-5 глаголов, в которых допускают ошибки (аргументировать, думать, доказывать…).
6) +частотники топ-частей речи по log-likelihood сравнить с референсным корпусом (НКРЯ), чтобы получить, напр., глаголы, специфичные для научной речи.
7) Можно рассматривать деепричастия: если явно не получается найти контролёра, то можно указать на рассогласование. ???

Задачи на ближайшее время:

- **разметка корпусов;**
- **Частотность: частотные списки по частям речи; читать про коллокационные меры: LL для частей речи, сравнение с референц-корпусом, TTR**
- **word2vec для определения тематики текста и поиска “синонимов”**
- **“Типичные” ошибки (синтаксис прежде всего). Формализация поиска типичных ошибок (сравн. констр., субъект деепричастной клаузы и т.д.).**

Прототип системы:

1. На вход подается студ. текст, который сравнивается: а) с большим референтным корпусом (reference corpus), б) с тематическим корпусом в том же домене (возможно, word2vec, чтобы определить, из какого домена текст).
2. Сравнение с реф. корпусом:
    a. Повторяющиеся леммы (топ-частот по частям речи, отличающиеся от распределения по большому корпусу)
    b. морф. и синт. ошибки: (несуществ. формы “столи”, ошибки согласования “большая стол”).

3. Сравнение с тематическим корпусом:
    a. Неспецифически научные слова и коллокации (топ-частот по частями речи и топ-частот биграмм и триграмм, “гапаксы” (слова уникальные для проверяемого текста)
    b. термины (как определять?)
4. Результаты
    a. Часть результатов относится ко всему текст: TTR, близость к домену X. (таблица в шапке анализа)
    b. Большая часть результатов подсвечивается в тексте:
        i. С вариантами замены:
        — Нестандартная синт. связь “большая стол”: вариант стандартной замены
        — С вариантами замены: нестандартное слово — синонимы по word2vec’у — коллокации по LL для этих слов.
        ii. Без вариантов замены:
        — “Этот слово редко встречается в текстах на эту тему. Убедитесь, что оно употреблено правильно.”

Тестирование: оценка асессорами. Можно дать 10 грязных текстов (~500 слов) асессорам (2-3-5), они размечают ошибки и затем мы сравниваем с результатом нашей программы. Из результатов асессоров нужно оставлять не 1 вариант, а несколько. Одна из мер качеств: точность - ложный alarm/точный; полнота - на все ли выделенные реакции человека есть реакция программы.