Диахронические эмбеддинги

Дарья Бакшандаева
Светлана Выдрина
Юлия Родина
Вадим Фомин
Куратор: Андрей Кутузов

Цель

Разработать систему, которая будет анализировать **русскоязычные новостные тексты** и выявлять семантические сдвиги определенных абстрактных понятий в рамках коротких промежутков времени* с помощью **дистрибутивной семантики.**

** недели, месяцы, годы*

“Война это мир, свобода это рабство, незнание - это сила” (Дж. Оруэлл)

Цель

В идеале:

1. **веб-сервис**, в реальном времени отслеживающий семантические сдвиги в русских СМИ;
2. **оценка** качества созданных моделей;
3. анализ **корреляции изменений с событиями реального мира**.

Минимум:

Масштабировать какой-либо существующий алгоритм на русский язык.

Задачи

- Проанализировать существующие алгоритмы,
- Определиться с датасетами,
- Определиться с исследуемыми понятиями и временными промежутками,
- Разработать / масштабировать модели,
- Применить модели для данных русского языка
- Найти способ оценки моделей,
- Оценить их,
- Интерпретировать результаты.

Материал

- **Газетный подкорпус НКРЯ** (НО: не очень новый и не очень большой),
- **Тайга** (хорошо, НО: не совсем полно и biased: ограничено кол-во источников)
- **Лента.ру** (хорошо, НО: biased, т. к. это один сайт)
- ....

Что ещё добавить?

Возможно, стоит определить какие-то критерии выбора новостных сайтов.

Related work

- Field A. et al. Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies // EMNLP-2018.
  - Упоминания США в “известиях” коррелируют с плохим состоянием экономики;
  - Можно исследовать, как состояние экономики связано с языком в “Известиях”
- Kutuzov A. et al. Diachronic word embeddings and semantic shifts: a survey // COLING-2018.
- Rosenfeld and Erk. Deep Neural Models of Semantic Shifts // NAACL-2018

Трудности и риски

- Нет нормального способа оценки.
- Как анализировать динамичный источник?
  - Если гранулярность 1 месяц, то нужно мин. 10 млн слов в месяц.
  - Нужны устойчивые каналы регулярного получения данных.
- Собственно перенос алгоритмов на русский язык и разработка.

Статья (предположительно на “Диалог”)

- Обзор литературы
- Архитектура
- Дизайн экспериментов
- Результат экспериментов
- Анализ связи между языковыми и общественными изменениями

Спасибо за внимание!