Выравнивание текстов в параллельном корпусе:

Lingtrain_Aligner

Темы курса

- Различные параллельные корпуса в составе НКРЯ (сложности создания)
- Alignment и как его делать
- Разметка
- Методы анализа, исследования на материале корпусов
- Построение датасета на примере пассива
- Извлечение контекстов из корпуса
- Семантические карты
- Графы
- Case study: Проект по церковно-славяским текстам
- Case study: Проект по текстам Нового Завета

Рекап

Основные понятия с прошлой пары:

1. Параллельные корпусы и примеры
2. Единица выравнивания
3. Основные типы выравнивания
4. Дистрибутивная семантика и векторные модели

Дистрибутивная семантика

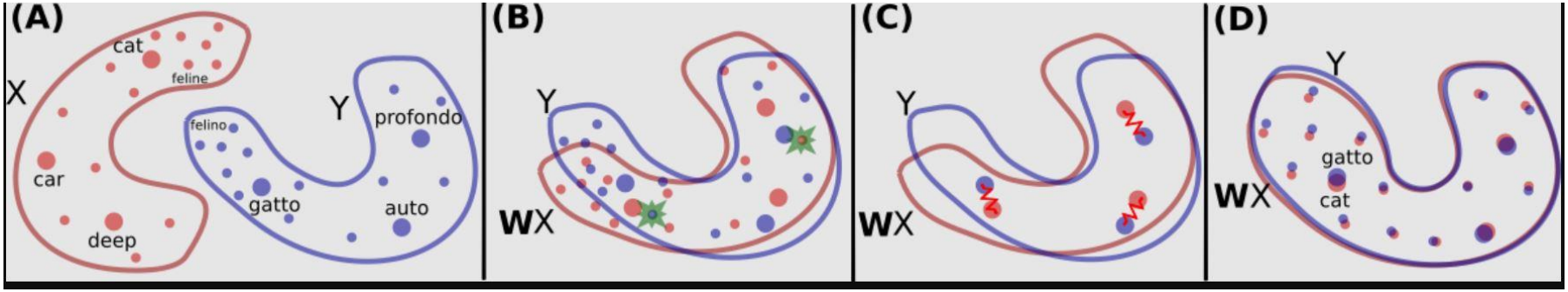
- Лекция Андрея Кутузова о математических основах векторных моделей
- RusVectōrēs: семантические модели для русского языка
- ShiftRy: is a web service for analyzing diachronic changes in the usage of words occurring in news texts from Russian mass media
- WebVectors: word embeddings online

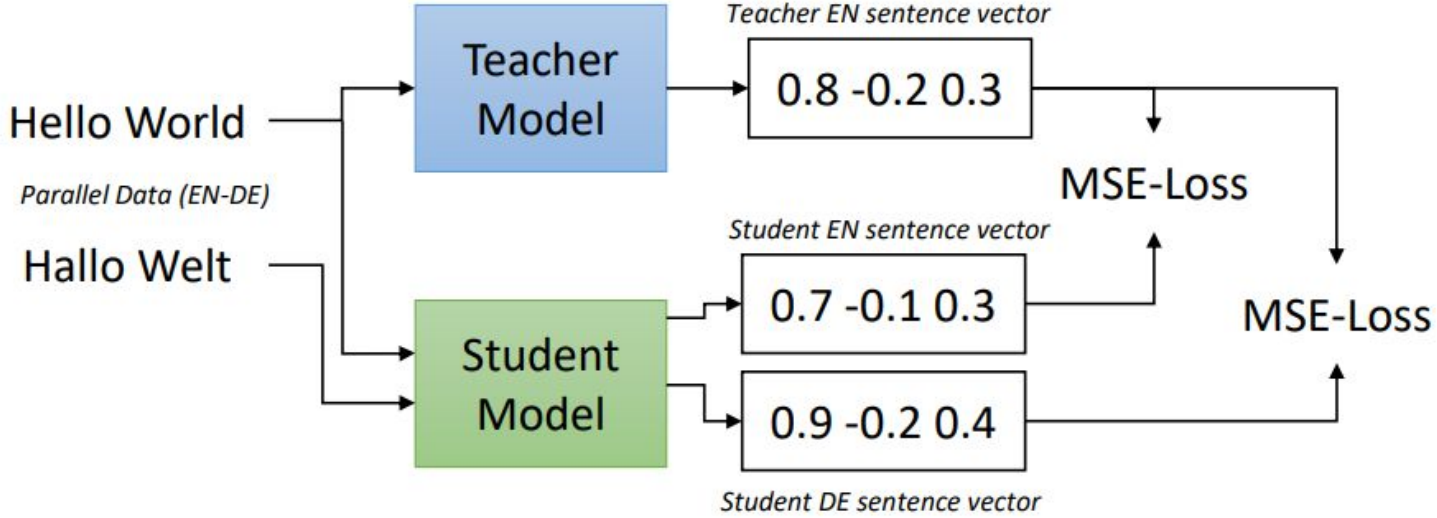
Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

Nils Reimers, Iryna Gurevych

We present an easy and efficient method to extend existing sentence embedding models to new languages. This allows to create multilingual versions from previously monolingual models. The training is based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence. We use the original (monolingual) model to generate sentence embeddings for the source language and then train a new system on translated sentences to mimic the original model

https://arxiv.org/pdf/2004.09813.pdf


(A)
cat
feline
X
Y
profondo
felino
car
gatto
auto
deep
(B)
Y
WX
(C)
Y
WX
(D)
Y
gatto
WX
cat


Teacher EN sentence vector
Teacher Model
0.8 -0.2 0.3
Hello World
MSE-Loss
Parallel Data (EN-DE)
Student EN sentence vector
Hallo Welt
0.7 -0.1 0.3
MSE-Loss
Student Model
0.9 -0.2 0.4
Student DE sentence vector

Посмотрим в код:)

```
russian = ["Кожа у него была свежая, совсем без пор, какая бывает у обжор-богачей.", #1#1
"Он оказался обжорой, заказал и закуски, и суп, и два вторых блюда.", #2#3
"Дон Кроче кивнул, на крупном лице цвета красного дерева царило сонное удовлетворенное выражение обжоры."]#3#2

english = ["His skin was fresh looking and poreless in the way of wealthy overeaters.",
"Don Croce nodded; his massive mahogany face wore the sleepy amiable look of the obese.",
"He turned out to be quite a glutton, ordering hors d'oeuvres, and soup, and two main dishes."]
```

Задание:

- откроем тетрадку
- в параллельном корпусе НКРЯ осуществить поиск по лексеме (любой) для 2 языков, выбрать 3-4 предложения, в которых она содержится
- заполнить списки в коде вашими предложениями и посмотреть на значения векторов

Lingtrain_Aligner

Данные: ParTY

Инструмент: Lingtrain Aligner

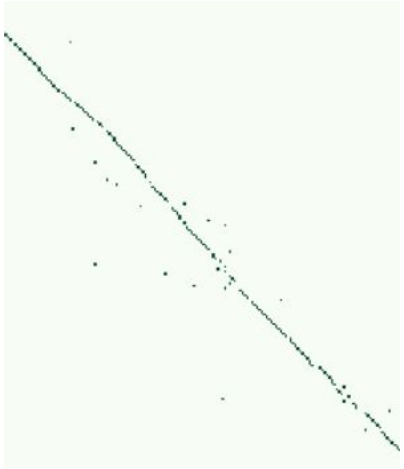
Что в основе? → Векторные модели, дистрибутивная семантика

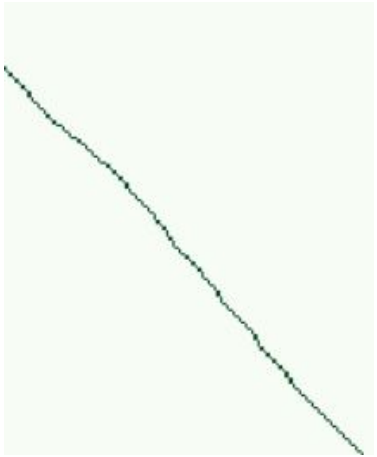
Lingtrain_Aligner

1. Найти тексты для выравнивания
2. Разметка

Token	Purpose	Mode
%%%%%title.	Title	Manual
%%%%%author.	Author	Manual
%%%%%h1. %%%%%h2. %%%%%h3. %%%%%h4. %%%%%h5.	Headings	Manual
%%%%%qtext.	Quote	Manual
%%%%%qname.	Text under the quote	Manual
%%%%%image.	Image	Manual
%%%%%translator.	Переводчик	Manual
%%%%%divider.	Divider	Manual
%%%%%.	New paragraph	Auto

Lingtrain_Aligner





1. Найти тексты для выравнивания
2. Разметка
3. Первичное выравнивание
4. Разрешение конфликтов:
   - слишком много хороших вариантов перевода одного предложения
   - нужное предложение не попала в окно выравнивания
   - The most frequent conflicts are the size of '2:3' and '3:2'. It means that one of the sentences here was translated as two or vise versa.
5. Немного подробнее https://habr.com/ru/post/586574/

Lingtrain_Aligner

- Colab_Version
- Потренируемся вместе:
  - тексты
  - тетрадка
- Самостоятельное выравнивание:
  - тексты (выбрать кусочек - можно около 50 реплик)
  - тетрадка та же
  - результаты выложить в папку, добавив фамилию в название html-файла (сегодня или в течение недели - 1 балл)

Не забываем про…

дневник :)