Выравнивание текстов в параллельном корпусе: подходы и инструменты

Рекап

1. Определение параллельного корпуса
2. Переводные единицы

Рекап

1. Определение параллельного корпуса
2. Переводные единицы

eng And here is the joke. A day earlier, she'd called him to ask for my phone number. [André Aciman. Find Me (2019) | A
снята] ←...→

rus И вот что забавно: **днем** ранее она связалась с ним, чтобы узнать мой номер. [André Aciman. Find Me (2019) |
снята] ←...→

3. Переводные эквиваленты

Рекап

1. Определение параллельного корпуса
2. Переводные единицы

eng | And here is the joke. A day earlier, she'd called him to ask for my phone number. [André Aciman. Find Me (2019) | A
снята] ←...→

rus | И вот что забавно: **днем** ранее она связалась с ним, чтобы узнать мой номер. [André Aciman. Find Me (2019) |
снята] ←...→

3. Переводные эквиваленты и нетривиальные соответствия в переводе
4. Pipeline создания параллельного корпуса

Примеры параллельных текстов

Примеры параллельных текстов

Наиболее известные:

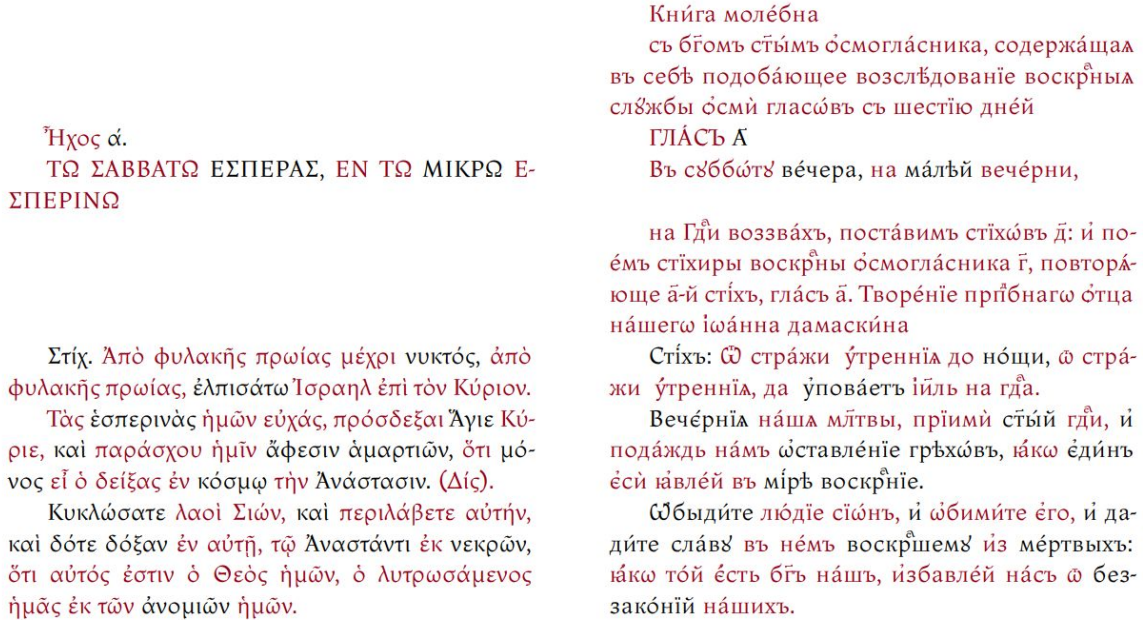
- Библия
- Гарри Поттер
- Маленький принц
- Винни Пух

Чуть менее тривиальные:

- Инструкции
- Субтитры к фильмам

Примеры параллельных корпусов

- Параллельный корпус для Слова о полку Игореве
- Выровненный программными средствами текст октоиха



Ἦχος α΄.

ΤΩ ΣΑΒΒΑΤΩ ΕΣΠΕΡΑΣ, ΕΝ ΤΩ ΜΙΚΡΩ ΕΣΠΕΡΙΝΩ

Στίχ. Ἀπὸ φυλακῆς πρωίας μέχρι νυκτός, ἀπὸ φυλακῆς πρωίας, ἐλπισάτω Ἰσραηλ ἐπὶ τὸν Κύριον.

Τὰς ἑσπερινὰς ἡμῶν εὐχάς, πρόσδεξαι Ἅγιε Κύριε, καὶ παράσχου ἡμῖν ἄφεσιν ἁμαρτιῶν, ὅτι μόνος εἶ ὁ δείξας ἐν κόσμῳ τὴν Ἀνάστασιν. (Δίς).

Κυκλώσατε λαοὶ Σιών, καὶ περιλάβετε αὐτήν, καὶ δότε δόξαν ἐν αὐτῇ, τῷ Ἀναστάντι ἐκ νεκρῶν, ὅτι αὐτός ἐστιν ὁ Θεὸς ἡμῶν, ὁ λυτρωσάμενος ἡμᾶς ἐκ τῶν ἀνομιῶν ἡμῶν.

Кни́га моле́бна

съ бг҃омъ ст҃ы́мъ ѻ҆смогла́сника, содержа́щаѧ въ себѣ подоба́ющее возслѣ́дованїе воскр҃ныѧ слв҃жбы ѻ҆смѝ гласѡ́въ съ шестїю дне́й

ГЛА́СЪ А҃

Въ сѵ҆ббѡ́тѵ҆ ве́чера, на ма́лѣй вече́рни,

на Гд҃и воззва́хъ, поста́вимъ стїхѡ́въ д҃: и҆ пое́мъ стїхиры воскр҃ны ѻ҆смогла́сника г҃, повторя́юще а҃-й стїхъ, гла́съ а҃. Твореніе прпбнагѡ ѻ҆тца на́шегѡ і҆ѡа́нна дамаскѝна

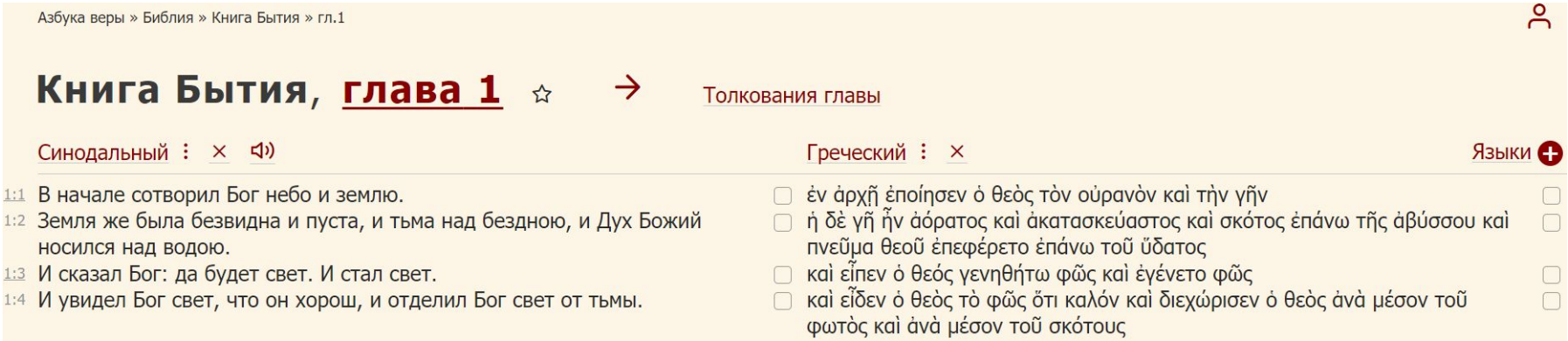
Стїхъ: Ѡ҆ стра́жи ѹ҆тренніѧ до но́щи, ѡ҆ стра́жи ѹ҆тренніѧ, да ѹ҆пова́етъ і҆и҃ль на гд҃а.

Вече́рніѧ на́шѧ мл҃твы, прїимѝ ст҃ы́й гд҃и, и҆ пода́ждь на́мъ ѡ҆ставле́ніе грѣхѡ́въ, ꙗ́кѡ е҆ди́нъ е҆сѝ ꙗ҆вле́й въ мі́рѣ воскр҃нїе.

Ѡ҆быди́те лю́дїе сїѡ́нъ, и҆ ѡ҆бими́те е҆го́, и҆ даи́те сла́вѹ҆ въ не́мъ воскр҃шемѹ҆ и҆з ме́ртвыхъ: ꙗ́кѡ то́й е́сть бг҃ъ на́шъ, и҆збавле́й на́съ ѡ҆ безако́нїй на́шихъ.

Примеры корпусов

1. Коллекция OPUS - платформа для работы с большим количеством языков, которая предлагает множество различных инструментов
2. ParTY (субтитры к фильмам)
3. Параллельный корпус в составе НКРЯ (художественные произведения)
4. Библия



Примеры корпусов

Корпус слушаний Европарламента

- 21 официальный язык ЕС
- Все подкорпуса выровнены по английскому
- XML, размечены говорящие, файл соответствует дню слушаний
- Свободный для скачивания

Корпус европейского права (The JRC-Acquis Multilingual Parallel Corpus)

- Действующее право ЕС. 22 языка
- Общий объём – 1 млрд слов
- Автоматическое выравнивание (венгерская программа HunAlign).

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

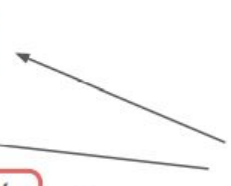
Выравнивание

Установление соответствия между единицами (абзацем, предложениями, словами) оригинального и переводного текста.

Основные подходы	Проблемы
→ пословное выравнивание	несовпадение лексем в разных языках
→ выравнивание по предложениям	количество предложений и их расположение в тексте может не совпадать (перекрестная структура текста)

Выравнивание по предложениях vs По словам

EN: He was interrupted by a knock on the door.
RU: В дверь постучали.
DE: Ein Klopfen an der Tür unterbrach ihn.
ITA: Fu interrotto da qualcuno *che bussava alla porta*.
ES: Le interrumpieron unos golpes en la puerta.
CZ: Přerušilo ho zaklepání na dveře.
BG: Прекъсна го почукване на вратата.
SE: Han avbröts av en knackning på dörren.
FR: Il fut interrompu par des coups frappés à la porte.



Переводные эквиваленты

Выравнивание по словам - основной принцип

Предложение языка L1: $s_{(1..j)} = s_1 + \ldots + s_j$ s → sourse

Предложение языка L2: $t_{(1..i)} = t_1 + \ldots + t_i$ t → target

Под выравниванием понимается отображение множества позиций слов исходного предложения {1 … j} во множество позиций слов целевого {1 … i} предложения:

$$a: i \rightarrow j$$

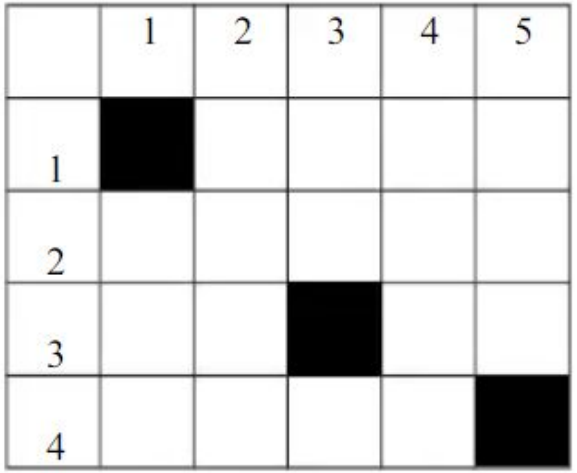
Выравнивание по словам

Пример пословного выравнивания с помощью программы Giza++:

Ls: Штифт (1) для (2) использования (3) в (4) стоматологии (5).

Lt: NULL(0) Implant(1) à (2) usage(3) dentaire(4).

a: { 1 → 1, 2 → 0, 3 → 3, 4 → 0, 4 → 5}



Потренируемся :)

Ls: Он(1) будет(2) рад(3) цветам(4).

Lt: NULL(0) He(1) will(2) love(3) the(4) flowers(5).

Ls: Луна(1) освещала(2) голые(3) обугленные(4) стволы(5).

Lt: NULL(0) The(1) moon(2) illuminated(3) bare(4), charred(5) tree(6) trunks(7).

Ls: У(1) меня(2) красивые(3) руки(4).

Lt: NULL(0) I(1) have(2) nice(3) hands(4).

Основные подходы к пословному выравниванию

1. Эвристические модели: для каждой пары соответствующих друг другу предложений строится матрица со значением ассоциативных мер между каждым словом исходного предложения и каждым словом целевого предложения. Для выравнивания со словом исходного предложения выбирается слово с большей мерой.

2. Самообучающиеся модели: модели, основанные на принципах вероятности (в их основе - машинное обучение и скрытые марковские цепи)

Выравнивание по словам - литература:

- Östling R., Tiedemann J. Efficient Word Alignment with Markov Chain Monte Carlo // The Prague Bulletin of Mathematical Linguistics. 2016. No 1 (106). С. 125–146
- Östling R. Word order typology through multilingual word alignment, In: The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: Proceedings of the Conference, Volume 2: Short Papers, 2015, p. 205-211

Выравнивание по предложениям

В наиболее простых случаях для деления текста на предложения используются элементарные синтаксические признаки конца и начала предложения – пунктуация, использование заглавных букв, знаки абзаца.

Основные подходы:

1. По длине (LF-aligner, Евклид - оболочки на базе HunAlign)
2. По лексике (алгоритмы дистрибутивной семантики, BERT, Lingtrain)

HunAlign. Выравнивание по предложениям

In the absence of a dictionary, it first falls back to sentence-length information, and then builds an automatic dictionary based on this alignment. Then it realigns the text in a second pass, using the automatic dictionary.

Like most sentence aligners, hunalign does not deal with changes of sentence order: it is unable to come up with crossing alignments, i.e., segments A and B in one language corresponding to segments B' A' in the other language.

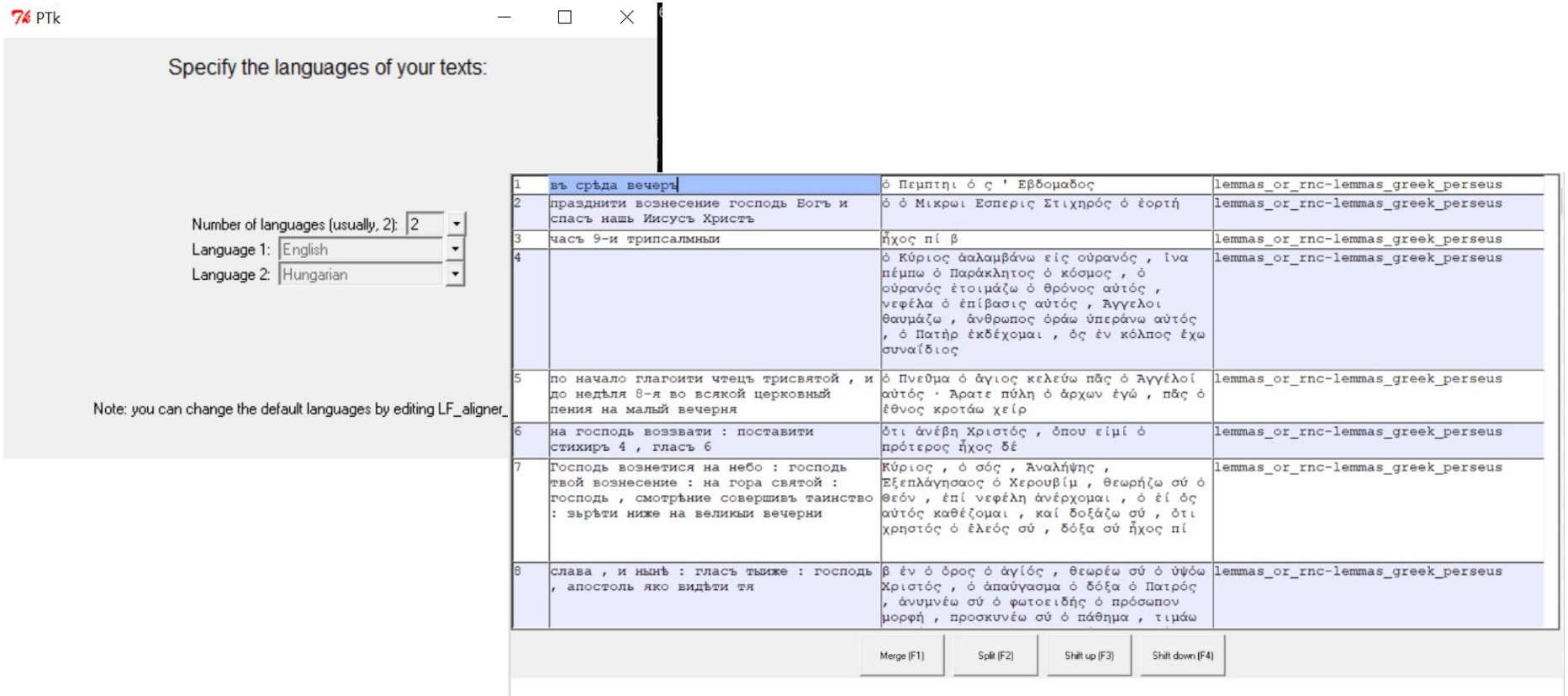
В HunAlign подгружаются лемматизированные файлы

Выравнивание по предложениям

«Евклид» (оболочка для HunAlign). Тексты выравниваются попарно и затем «склеиваются» в единый XML в соответствием с разделением предложений в оригинале:

```
<para id="2">
  <se lang="ru" variant_id="0">Марта 25 числа случилось в Петербурге необыкновенно странное происшествие.</se>
  <se lang="fr" variant_id="1">Le 25 mars, un événement tout à fait étrange s'est produit à Pétersbourg.</se>
  <se lang="fr" variant_id="2">Ce jour-là, 25 mars dernier, Pétersbourg fut le théâtre d’une aventure des plus étranges.</se>
  <se lang="fr" variant_id="3">Le 25 mars il est arrivé à Pétersbourg un événement extrêmement bizarre.</se>
</para>
```

Выравнивание по предложениям. LF-aligner



Дистрибутивная семантика

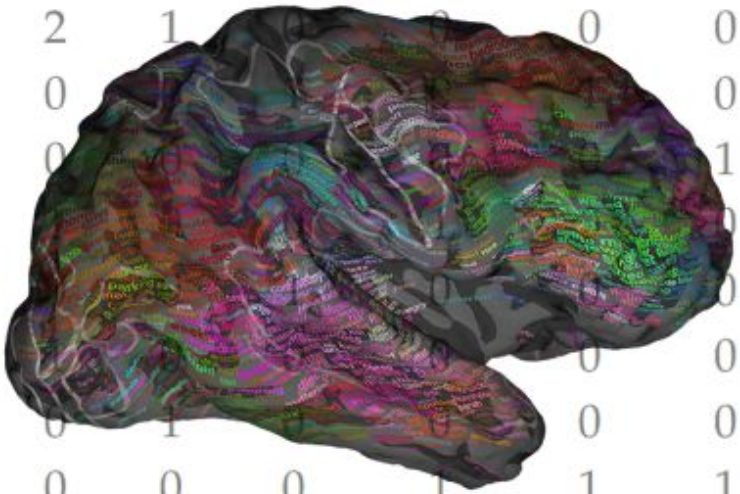
Дистрибутивная семантика - это метод исследования языка, основанный на дистрибуции отдельных единиц в тексте и не использующий сведений о конкретном лексическом и грамматическом значении слов.

Впервые предложен Л.Блумфильдом в 20ые гг. XX века и применялся, главным образом, в фонологии и морфологии. Используя контексты в качестве исходных данных (какие единицы могут взаимодействовать друг с другом), можно выделить основные единицы языка, объединить в классы и установить отношения встречаемости между ними.

Дистрибутивная семантика позволяет автоматически устанавливать контексты употребления слова и устанавливать семантические связи между лексемами.

Дистрибутивная семантика

	I	*like*	*enjoy*	*deep*	*learning*	*NLP*	*flying*	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	0	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	0	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0



We want a machine to imitate human brain and understand meaning of words.

Дистрибутивная семантика

В качестве инструмента дистрибутивной семантики используется линейная алгебра. Информация о дистрибуции единиц представляется в виде многоразрядных векторов.

Сами векторы соответствуют лексическим единицам (словам или словосочетаниям), а измерения векторов - контекстам, в которых эти единицы встречаются.

$$A_{m,n} = \begin{array}{c} \\ \textit{coffee} \\ w_2 \\ \textit{tea} \\ \vdots \\ w_m \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & \textit{drink} & w_3 & \dots & w_m \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & \dots & 1 \\ 1 & 0 & 2 & \dots & 0 \\ 0 & 2 & 0 & \dots & 3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \dots & 0 \end{bmatrix} \end{array}$$

Словесное пространство для слов английского языка *coffee* и *tea*

Дистрибутивная семантика

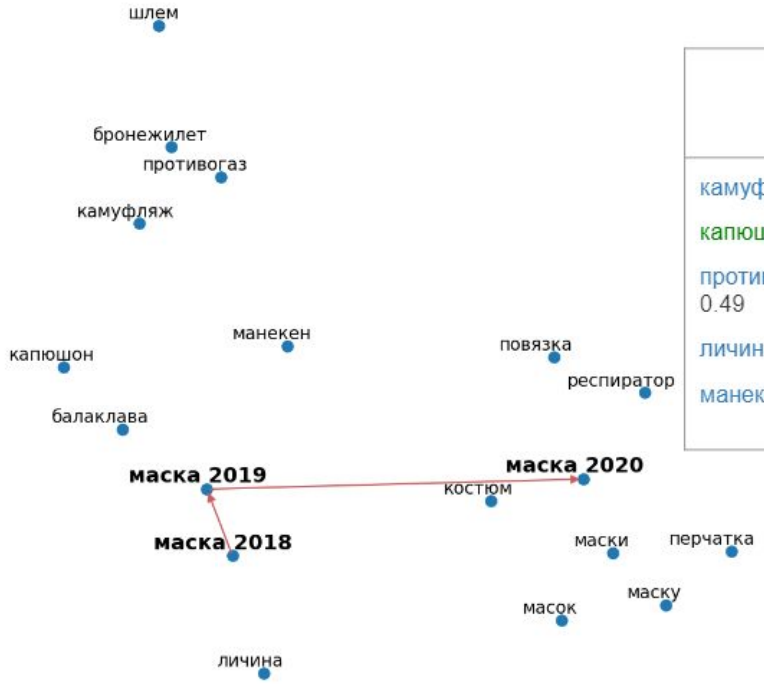
Семантическая близость между единицами вычисляется как косинусное расстояние между векторами.

$$cos(w1, w2) = \frac{\vec{V}(w1) \times \vec{V}(w2)}{|\vec{V}(w1)| \times |\vec{V}(w2)|}$$

$$cos(tomat, philosophy) = 0.00698$$
$$cos(pomidor, philosophy) = -0.03429$$
$$cos(tomat, pomidor) = 0.65049$$

https://www.hse.ru/data/2016/11/02/1111119007/ismw.pdf

Дистрибутивная семантика



2018	2019	2020 Есть сдвиг! (0.64)
камуфляж NOUN 0.52 капюшон NOUN 0.51 противогаз NOUN 0.49 личина NOUN 0.47 манекен NOUN 0.45	камуфляж NOUN 0.57 капюшон NOUN 0.50 перчатка NOUN 0.47 шлем NOUN 0.45 противогаз NOUN 0.45	повязка NOUN 0.60 одноразовый ADJ 0.55 костюм NOUN 0.50 защитный ADJ 0.49 перчатка NOUN 0.46

https://shiftry.rusvectores.org/en/about/

Дистрибутивная семантика

С помощью дистрибутивной семантики легко и быстро решается множество проблем: снятие семантической неоднозначности, тематическая кластеризация, генерация тезаурусов, поиск синонимов, антонимов, гиперонимов и многое другое.

Чем полезно для выравнивания по предложениям?

Дистрибутивная семантика и выравнивание текстов

Lingtrain Aligner: https://github.com/averkij/lingtrain-aligner

В чем идея:

→ Модель берет заданное количество строк первого текста и подбирает в соответствующем фрагменте второго текста лучшие соответствия, используя векторные представления.

[Ὁ Οἶκος]

[Τὸν πρὸ ἡλίου Ἥλιον, δύναντα ποτὲ ἐν τάφῳ, προέφθασαν πρὸς ὄρθρον, ἐκζητοῦσαι ὡς ἡμέραν,
Μυροφόροι κόραι, καὶ πρὸς ἀλλήλας ἐβόων· Ὦ φίλαι, δεῦτε τοῖς ἀρώμασιν ὑπαλείψωμεν,
Σῶμα ζωηφόρον καὶ τεθαμμένον, σάρκα ἀνιστῶσαν τὸν παραπεσόντα Ἀδὰμ κείμενον ἐν τῷ μνήματι, ἄγωμεν,
σπεύσωμεν, ὥσπερ οἱ Μάγοι, καὶ προσκυνήσωμεν, καὶ προσκομίσωμεν τὰ μύρα ὡς δῶρα τῷ μὴ ἐν σπαργάνοις,
ἀλλ' ἐν σινδόνι ἐνειλημένῳ, καὶ κλαύσωμεν, καὶ κράξωμεν· Ὦ Δέσποτα ἐξεγέρθητι, ὁ τοῖς πεσοῦσι παρέχων ἀνάστασιν.]

[Συναξάριον]

[Τῇ ἁγίᾳ καὶ μεγάλῃ Κυριακῇ τοῦ Πάσχα, αὐτὴν τὴν ζωηφόρον Ἀνάστασιν ἑορτάζομεν τοῦ Κυρίου, καὶ Θεοῦ καὶ Σωτῆρος ἡμῶν Ἰησοῦ Χριστοῦ.]

[Στίχοι]

[Χριστὸς κατελθὼν πρὸς πύλην Ἅδου μόνος]

[Λαβὼν ἀνῆλθε πολλὰ τῆς νίκης σκῦλα.]

[Ико́с:]

[Еже́ пре́жде со́лнца Со́лнце заше́дшее иногда́ во гроб, предвари́ша ко у́тру, и́щущия я́ко дне мироно́сицы де́вы,
и друга́ ко друзе́й вопия́ху: о други́ни! Прииди́те, воня́ми пома́жем те́ло живоно́сное и погребе́ное,
плоть воскреси́вшаго па́дшаго Ада́ма, лежа́щую во гро́бе. Иде́м, потщи́мся я́коже волсви́, и поклони́мся, и принесе́м ми́ра я́ко да́ры,
не в пелена́х, но в плащани́це обви́тому, и пла́чим, и возопии́м: о Влады́ко, воста́ни, па́дшим подая́й воскресе́ние.]

[Синаксарь, во святу́ю и вели́кую Неде́лю Па́схи.]

[Стихи́: Христо́с сше́д к борьбе́ а́дове Еди́н.]

[Мно́гия взе́м побе́ды кормы́сти, взы́де.]

[Во святу́ю и вели́кую Неде́лю Па́схи, са́мое живоно́сное Воскресе́ние пра́зднуем Го́спода Бо́га и Спа́са на́шего Иису́са Христа́]

CSL	GR	1	2	3	4	5	6	7
1	[[	0.393678	0.04205396	-0.24479382	-0.05323773	-0.24479382	0.0911667	-0.01963074]
2	[	-0.01602886	0.34094387	-0.13031456	0.22283232	-0.13031456	0.1565712	0.11568692]
3	[	0.07481634	0.21875925	-0.17499971	0.5257409	-0.17499971	0.23422936	-0.05796702]
4	[	0.2000877	0.11701456	-0.23621589	0.16487566	-0.23621589	0.26789212	0.02254544]
5	[	0.00616331	0.0261609	0.08375084	-0.0773346	0.08375084	0.03489467	0.359214]
6	[	0.03978096	0.32326674	-0.27889502	0.7659529	-0.27889502	0.27096146	0.00304512]]

3. [Συναξάριον]

4. [Τῇ ἁγίᾳ καὶ μεγάλῃ Κυριακῇ τοῦ Πάσχα, αὐτὴν τὴν ζωηφόρον Ἀνάστασιν ἑορτάζομεν τοῦ Κυρίου, καὶ Θεοῦ καὶ Σωτῆρος ἡμῶν Ἰησοῦ Χριστοῦ.]

5. [Στίχοι]

6. [Χριστὸς κατελθὼν πρὸς πύλην Ἅδου μόνος]

7. [Λαβὼν ἀνῆλθε πολλὰ τῆς νίκης σκῦλα.]

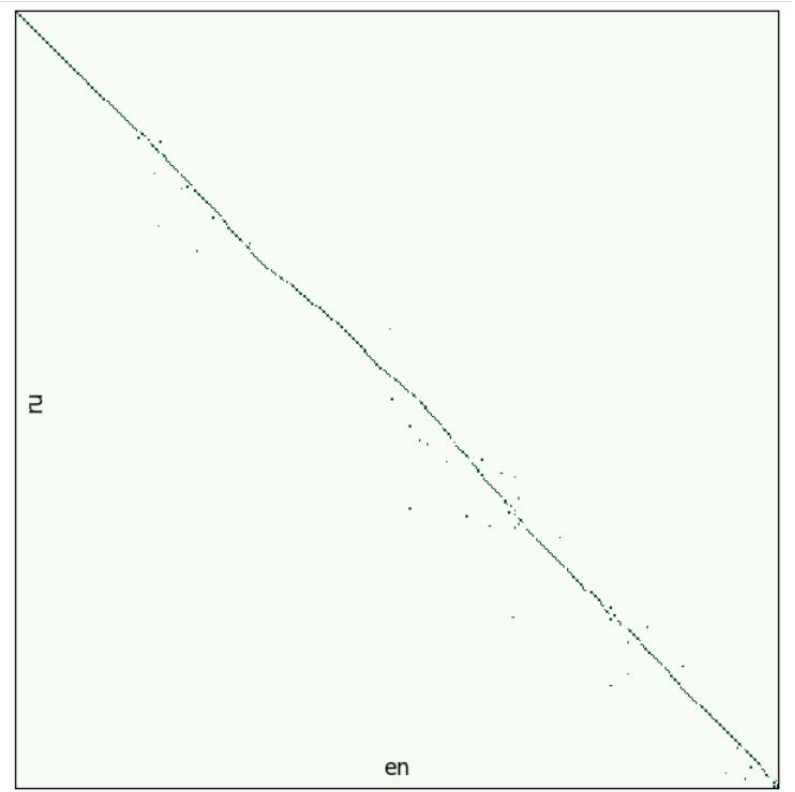
3. [Синаксáрь, во святýю и велúкую Недéлю Пáсхи.]

4. [Стихи́: Христóс сшéд к борьбé áдове Едúн.]

5. [Мнóгия взéм побéды коры́сти, взы́де.]

6. [Во святýю и велúкую Недéлю Пáсхи, сáмое живонóсное Воскресéние прáзднуем Гóспода Бóга и Спáса нáшего Иисýса Христá]

Дистрибутивная семантика и выравнивание текстов



Спасибо за внимание!