

# Творческая работа: выравнивание текстов

Л. Шляхтина

8 декабря 2024 г.

# Введение

- ▶ Задача: создать параллельный корпус из девятой главы "Гарри Поттер и Философский камень" на русском, японском и английском языках для использования в курсовых/дипломных работах.
- ▶ Проблема: небуквальный перевод, неудобно искать по какому-то одному слову.

## Этапы

Разбить текст на предложения, прогнать их через нейросеть, получить вектора и пары предложений по "похожести".

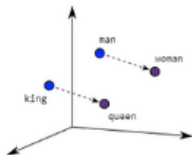
# Парсинг

С помощью библиотек `razdel` (часть проекта `Natasha`) и `hanami` разделяем имеющийся текст на предложения.

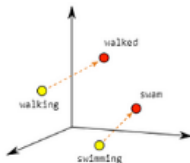
Язык	Кол-во предложений
Английский	388
Русский	490
Японский	317

Word2vec на вход принимает большой корпус и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Слова, встречающиеся в тексте рядом с одинаковыми словами, будут иметь близкие по косинусному расстоянию векторы. Векторные представления позволяют вычислять «семантическое расстояние» между словами и находить похожие по значению слова.

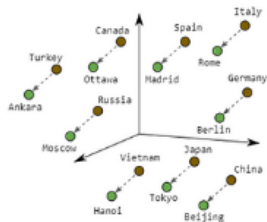
## Word2Vec



Male-Female




Verb Tense



Country-Capital

# Как это работает?

 Sentence Similarity Examples ▾

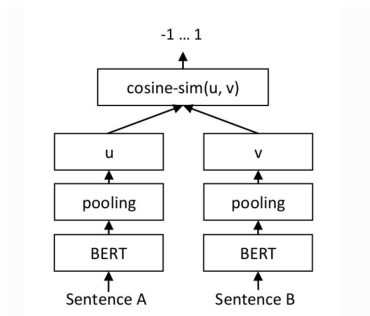
Source Sentence

Sentences to compare to

Computation time on cpu: cached

<div><div></div></div> That is a happy dog	0.833
<div><div></div></div> That is a very happy person	0.985
<div><div></div></div> Today is a sunny day	0.755

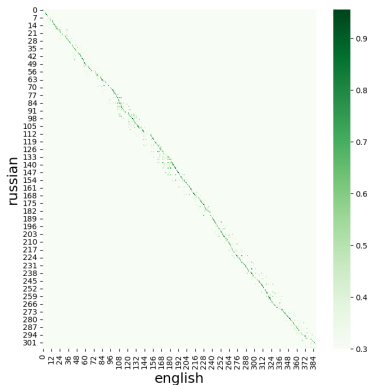
(a) Похожесть предложений



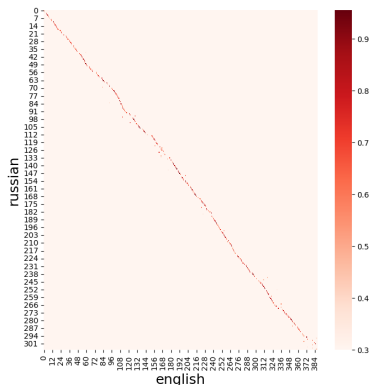
(b) Сравнение предложений

Рис. 1: Эмбединги предложений

# Визуализация матрицы



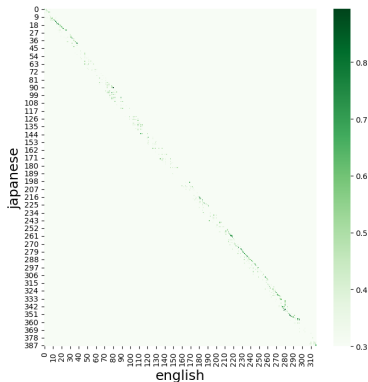
(a) Изначальная матрица



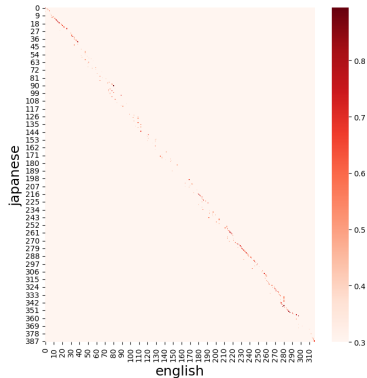
(b) Наиболее близкий вариант

Рис. 2: Матрица для русского и английского текстов

# Визуализация матрицы



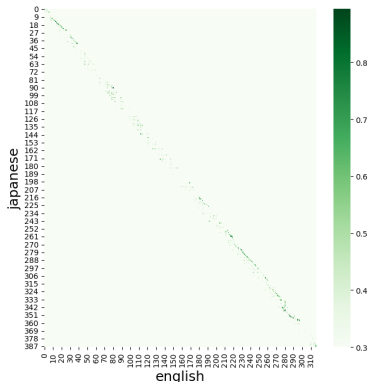
(a) Изначальная матрица



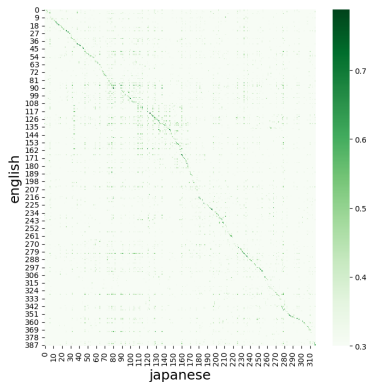
(b) Наиболее близкий вариант

Рис. 3: Матрица для японского и английского текстов

# Визуализация матрицы



(a) Матрица Sentence Transformers



(b) Матрица Universal Sentence Encoder

Рис. 4: Матрица для японского и английского текстов



# Результаты

	A	B	C	D	E	F
1		ru	en	sim_x	ja	sim_y
2	0	До приезда в Хогвартс Гарри и	Harry had never believed he would meet	0,65313	ダドリーより嫌なヤツがこの世の中	0,661952
3	1	Но это было до того, как он вс	Still, first-year Gryffindors only had Potio	0,432518	一年生ではグリフィンドールとスリ	0,502783
4	2	Однако, вернувшись от Харпи	Or at least, they didn't until they spotted	0,433398	少なくとも、グリフィンドールの談	0,512457
5	3	Со вторника начинались поле	Flying lessons would be starting on Thurs	0,805359	ー飛行訓練は木曜日に始まりです	0,599966
6	4	— Великолепно, — мрачно за	"Typical," said Harry darkly.	0,824407	ダドリーより嫌なヤツがこの世の中	0,396301
7	5	— Как Раз то, о чем я всегда м	"Just what I always wanted.	0,625262		
8	6	Выставить себя дураком пере	To make a fool of myself on a broomstick	0,471104		
9	7	Нужно отметить, что до того ка	He had been looking forward to learning t	0,40806	グリフィンドールとスリザリンとの	0,332261
10	8	— Откуда ты знаешь, кто буде	"You don't know that you'll make a fool o	0,678392	「そうなるとはかぎらないよ。あい	0,407553
11	9	— Разумеется, я знаю, что Ма	"Anyway, I know Malfoy's always going o	0,628951	「そうなるとはかぎらないよ。あい	0,503017
12	10	Малфой действительно через	Malfoy certainly did talk about flying a lot	0,893054	マルフォイの長ったらしい自慢話は	0,479667
13	11	Он во всеуслышание сожалел	He complained loudly about first years ne	0,64819	マルフォイは確かによく飛行の話を	0,601128
14	12	Впрочем, Малфой был не еди	He wasn't the only one, though: the way s	0,673113	シェーマス・フィネガンは、子供の	0,530501
15	13	Даже Рон готов был рассказать	Even Ron would tell anyone who'd listen	0,741261	ロンでさえ、聞いてくれる人がいれ	0,777102
16	14	Вообще все, что родился в се	Everyone from wizarding families talked	0,757941	魔法使いの家の子はみんなひっきり	0,708761
17	15	Из-за квиддича Рон уже успе	Ron had already had a big argument with	0,713867	ロンも同室のディーン・トーマスと	0,786272
18	16	Дин обожал футбол, а Рон ут	Ron couldn't see what was exciting about	0,684106	ロンにしてみれば、ボールがたった	0,727412
19	17	На следующий день Гарри за	Harry had caught Ron prodding Dean's po	0,649556	ディーンのお好きなウエストハム・サ	0,725183
20	18	Так, Невилл признался, что у	Neville had never been on a broomstick i	0,662805	ネビルは今まで一度も箒に乗ったこ	0,586743
21	19	В глубине души Гарри был с	Privately, Harry felt she'd had good reaso	0,57369	だいたいネビルは両足が地面に着い	0,598513
22	20	Гермиона Грэйнджер, как и Г	Hermione Granger was almost as nervous	0,744123	ハーマイオニー・グレンジャーも飛	0,629443
23	21	Если бы полетам можно был	This was something you couldn't learn by	0,339502	こればかりは、本を読んで暗記す	0,599999
24	22	Во вторник за завтраком она	At breakfast on Thursday she bored them	0,727837	木曜日の朝食の時ハーマイオニーは	0,63802
25	23	Но остальные были очень рад	Neville was hanging on to her every word	0,517863	ハグリッドの手紙の後、ハリーには	0,427127
26	24	С пятницы Гарри не получил	Harry hadn't had a single letter since Hagr	0,744303	ハグリッドの手紙の後、ハリーには	0,704205
27	25	Сова Малфоя — точнее, в отл	Malfoy's eagle owl was always bringing hi	0,585396		
28	26	В общем, Гарри во вторник не	A barn owl brought Neville a small packag	0,672112	ネビルは今まで一度も箒に乗ったこ	0,374422
29	27	Казалось, что шар заполнен	He opened it excitedly and showed them	0,555827		
30	28	— Это напоминалка! — поясн	"It's a Remembrall!" he explained.	0,717924		
31	29	— Бабушка знает, что я посто	"Gran knows I forget things — this tells y	0,705514	「『思いだし玉』だ!ばあちゃんは	0,452245
32	30	Вот смотрите — надо взять е	Look, you hold it tight like this and if it tur	0,688385	「『思いだし玉』だ!ばあちゃんは	0,39472