**Отчет по мастерской «Автоматический перевод анкет»**

**Эксперимент**

*Данные*:

* Русско-итальянский параллельный корпус (101814 предложений)
* Англо-русский параллельный корпус (527782 предложений)
* Англо-русский параллельный корпус «Гарри Поттер»

*Прилагательные на перевод:*

1. поле «smooth»:

* скользкий, ровный, плоский, гладкий, прямой (для русско-итальянского исследования)
* slick, level, flat, smooth, straight, slippery, greasy (для англо-русского исследования)

1. поле «sharp»:

* острый
* sharp, pointed

1. поле «size»:

* тонкий, толстый, широкий, узкий, тесный, просторный
* thick, thin, broad, narrow, wide, large, spacious

*Существительные на перевод:*

Подробнее в файле nouns\_for\_translation.xlsx

*Метод:*

Нахождение перевода с помощью частеречной разметки и выравненного текста (предложение в оригинале; перевод)

В эксперименте для выявления зависимости использовалась следующая метрика:

,

где P (A|B) - вероятность присутствия прилагательного A в выровненном целевом предложении с учетом прилагательного B в исходном предложении

*Результаты:*

Для вычисления, насколько метод хорошо сработан использовалась точность (precision) и полнота (recall). Золотой стандарт определялся носителями языка (представлен в файле results.xlsx)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Adjectives | | | | | |
|  |  | sharp | smooth | big | small |
| en-ru | Accuracy: | 1,000 | 0,636 | 0,429 | 0,667 |
| Completeness: | 1,000 | 1,000 | 1,000 | 0,667 |
| ru/en-it | Accuracy: | 1,000 | 0,500 | 0,600 | 0,500 |
| Completeness: | 0,167 | 0,091 | 0,375 | 0,250 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nouns | | | | |
|  |  | sharp | smooth | size |
| en-ru | Accuracy: | 0,393 | 0,417 | 0,434 |
| ru/en-it | Accuracy: | 0,500 | 0,396 | 0,404 |

**Идеи**

а) Нахождение кандидатов на перевод при помощи векторного представления слов

Смысл состоит в том, что слова со схожим значением имеют маленькое косинусное расстояние между векторами. Можно взять уже натренированные модели для двух языков, найти вектор нужного нам слова (которое нужно перевести) и сравнить косинусное расстояние между этим вектором и вектором каждого слова из предложения-перевода. Затем можно взять топ-5 кандидатов из каждого предложения и как-то их отсортировать. Например, применить частеречную разметку. Или, например, можно изначально взять все прилагательные / существительные из предложения-перевода и находить косинусную близость между ними и итогом будет уже топ-1.

б) Увеличение корпуса

Добавление ещё одного корпуса привело к увеличению показателей точности и полноты.