Отчет

Перед началом проверки гипотез мной был проведен анализ различных существующих подходов. Среди большого количества подходов мной были проверены следующие:

1) **BERT+DP (Amrami and Goldberg, 2019)** – лучший подход на данный момент для решения задачи WSI[1]. Несмотря на очень хорошее качество на английском языке, в данной задаче (с использованием **DeepPavlov/rubert-base-cased**) было получено качество **0.1 ARI**.
2) Методы, занявшие лидирующие места в исходном соревновании[2]:
   a. **1-е место**. Найти смыслы слова использую предобученные эмбеддинги и классифицировать предложения по косинусному расстоянию до них.
   b. **2-е место**. Кластеризовать взвешенное среднее эмбедингов контекста для получения смыслов[3].

   Лучшее качество получилось при использовании второго подхода **0.22 ARI** с использованием RusVectores эмбеддингов.
3) В рамках улучшения качества предыдущих подходов, мной была взята модель RusVectores **ELMo**[4] с контекстуальными эмбеддингами, но качество получилось хуже **0.14 ARI**.

Подход, который дал лучшее качество **0.232 ARI (трейн 0.169 ARI)**, заключался в **подготовке и чистке данных** для adagram. Были применены следующие приемы:

1) Удаление стоп-слов **+0.02 ARI**;
2) Удаление токенов, в которых содержаться цифры **+0.005 ARI**;
3) Удаленик токенов длинной меньше 2 (без этого остается много мусора в данных) **+0.003 ARI**;
4) Удаление токенов с большими буквами в них, кроме начала предложения (своеобразный NER) **+0.003 ARI**
5) Удаление английских слов (так как модель их не знает)
6) Использование **pymorphy2 POS-tags** для очистки нерелевантных тегов (оставляем только существительные, глаголы и прилагательные) **+0.022 ARI**;
7) Удаление дублирующихся токенов **+0.002 ARI**
8) ~~Удаление длинных слов~~ (не использовалось в итоговом алгоритме из-за возможного переобучения) **+0.005 ARI**

Если же удалить длинные слова и убрать инфинитивы глаголов, то на тестовых данных можно достигнуть качества **0.257 ARI**, но на тренировочных качество будет **0.137 ARI**, что говорит нам о том, что распределения данных в трейне и тесте отличаются (это видно, например, по длине предложений), но также, нельзя исключать возможное переобучение.

Лучший результат:

1) **0.232 ARI (трейн 0.169 ARI)**
2) **0.257 ARI (трейн 0.137 ARI)**

---

[1] http://nlpprogress.com/english/word_sense_disambiguation.html
[2] https://arxiv.org/pdf/1803.05795.pdf
[3] https://github.com/akutuzov/russian_wsi
[4] https://github.com/ltgoslo/simple_elmo