Дополнение к отчету по НИР весна 2020 10 семестр

Козлинский Евгений

1 Сделано

Поставил задачу притягивания карточек услуг с описаниями к рубрикатору. Критерием качества является полнота, то есть хотим притянуть как можно больше услуг к рубрикатору среди тех, которые можно притянуть. Точность гипотез по принадлежности услуги к рубрике стоит во втором приоритете.

Причем есть как описания, для которых место в рубрикаторе существует, так и те, которые не могут встроиться в нынешний рубрикатор. Таких услуг около половины. Но так как точность формирования гипотез стоит на втором месте, то мы не боремся с этой проблемой.

Для эксперимента я взял ветвь рубрикатора в 700 листьев, с суммарным количеством описаний услуг 850к, из них около 120к не имеют привязку к рубрикатору. Каждое описание представляет из себя от предложения до абзаца в 2 тысячи символов.

Есть базовая модель, она основана на пересечении слов в названии рубрики и описании услуги + синонимы. Эта модель покрывает малое количество услуг: полнота около 18%. Еще есть ряд моделей, основанных на близости векторных представлений (word2vec из коробки и емебединги из Bert), которые дали полноту не выше, чем 36%.

Я сделал модель на основе TM, которая справляется немного лучше, но еще нуждается в доработке. Рубрикатор фиксирован, поэтому чтобы сопоставить ему темы TM я использую не hARTM, а напрямую указываю в данных известные уровни рубрики в качестве модальностей (отдельная модальность на каждый уровень, их три). Настроил веса модальностей, чтобы модель обращала внимание как на текст, так и на уровни из модальностей + результат дал регуляризатор декоррелирования тем.

Для предсказания рубрик я использую два способа:

первый - предсказываю наиболее вероятные значения неизвестных модальностей

второй - для описания предсказываю наиболее вероятную тему, после чего определяю самую популярную (мода) рубрику из рубрикатора в этой теме. Моду считаю так: для всех документов обучающей выборки, для которых данная тема наиболее вероятна беру их рубрику. На полученном наборе рубрик ищу моду.

Обе модели дают неплохую полноту около 80% на тесте на искусственной выборке, созданной из набора описаний, про которые точно известно, что они занимают правильное место в рубрикаторе, путем удаления информации о местоположении в рубрикаторе.

На натуральной разметке (описания, которые изначально не были привязаны к рубрикатору, для которых человеческой разметкой были найдены места в рубрикаторе) первая модель дала полноту 37%, вторая модель дала полноту 44%.

2 Планы

- увеличить объем натуральной разметки. Сейчас в ней 120 примеров, хочется увеличить до 500.

- Начать делать множественный прогноз, и считать, что мы угадали рубрику, если хоть одни кандидат соответствует истинной рубрике.

- Попробовать обучить hARTM, гипотеза следующая: не нужно будет вешать большой вес на модальности, а значит модель будет больше полагаться на текст описания, что должно позитивно повлиять на предсказания на описаниях без модальностей.

- Построить модель на основе мер вложенности из бакалаврской работы Цыгановой Светланы 2013