Домашнее задание 3. Grouping Verbs to Frame Type Clusters

Deadline: 30.11.2018

Домашнее задание посвящено задаче определения смысла глаголов на данных соревнования SemEval 2019 Task 2 Subtask 1 (Grouping Verbs to Frame Type Clusters). Задание структурно похоже на задачу Word Sense Induction (WSI), рассмотренную на лекции, и предыдущие WSI соревнования, такие как RUSSE 2018. Однако в данном случае предлагается кластеризовать словоупотребления глаголов по их семантическим фреймам, а не словоупотребления существительных по их смыслам. Здесь инвентарь фреймов выступает в виде инвентаря смыслов.

Описание соревнования доступно по ссылке.
Регистрация необходима для получения данных.
Срипты для оценки доступны по ссылке.

Постановка задачи: даны глаголы и различные контексты их употребления. Смысл глагола определяется фреймом – скруктурой, которая представляет глагол как предикат с различными аргументами). Требуется так кластеризовать глаголы, чтобы каждый кластер соответствовал одному фрейму.

Этапы решение задачи:

1. Векторизация глаголов и предложений.
2. Кластеризация глаголов. Каждому употреблению глагола приписать метки кластеров: к одному кластеру относятся глаголы, относящиеся к одному фрейму.
3. Сравнение с золотым стандартом, основанном на FrameNet с использованием стандартных скриптов.

Вам предстоит реализовать три подхода к решению задачи: подход на основе эмбеддингов слов, эмбеддингов предложений и end-to-end. Вы можете использовать как предобученные модели, так и дообучать их на своих данных. Обязательно опишите все проведенные эксперименты в отчете. Каждый пункт вычислений должен быть прокомментирован и описан. Использование любых моделей должно быть покреплено описанием мотивации: зачем? почему? Оцените все подходы с помощью готовых скриптов для оценивания качества и укажите полученное качество в отчете.

Задание 1 (2 балла + количество моделей эмбеддингов)

1. Представление предложения: усредненный вектор эмбеддингов слов. Некоторые редобученные модели: ELMO, FastText, GloVe, word2vec.

2. Алгоритм кластеризации: произвольный.

Задание 2 (2 балла + количество моделей эмбеддингов)

1. Представление предложения: модель эмбеддинга предложения.
   Известные модели: StarSpace, Skip-thoughts, Sent2vec, USE, InferSent.
2. Алгоритм кластеризации: произвольный.

Задание 3 (3 балла)

1. Попробуйте совместить векторизацию предложений и кластеризацию в единую модель.
2. Оптимизируйте параметры модели.

Задание 4 (бонус, до 3 баллов)

1. Придумайте способ красиво и информативно визуализировать решение задачи.
2. Напишите небольшой обзор литературы по задаче WSI.

Рекомендуемое чтение

1. *Arefyev, Nikolay, Pavel Ermolaev, and Alexander Panchenko.* How much does a word weigh? Weighting word embeddings for word sense induction.
   https://arxiv.org/abs/1805.09209.
2. *Christian S. Perone, Roberto Silveira, Thomas S. Paula.* Evaluation of sentence embeddings in downstream and linguistic probing tasks:
   arxiv.org/abs/1806.06259
3. *Sanjeev Arora, Yingyu Liang, Tengyu Ma.* A simple but tough to beat baseline for sentence embeddings.
   https://openreview.net/pdf?id=SyK00v5xx