Advanced NLP (CS7.501)

Monsoon 2022, IIIT Hyderabad
Project Outline

Team Name: PyCrusts
Team Number: 18

Team Members:

- Jatin Agarwala (2020111011)
- Abhinav S Menon (2020114001)

Problem Statement

Problem Statement

We plan to implement a system that embeds words into a space of probability distributions, rather than simple vectors. What this entails is that distance in the space is factored by metrics other than cosine similarity – specifically, optimal transport.

Motivation

This approach is motivated primarily by concerns of interpretability. Word vectors, as we know, are uninterpretable (in the sense that we do not understand the correspondence between dimensions in the embedding space and semantic features). We also do not understand the distance metric that word vectors make use of (cosine similarity), in that the equation conveys no meaning to us as to why the model ranks two words as similar or dissimilar.

Treating words as distributions over contexts, however, allows us to interpret the model's view of words, and tweak it if necessary. Furthermore, the distance metric (optimal transport, borrowed from probability theory), by its nature gives us more information than a function like cosine similarity. It is of the form

$$\mathrm{OT}(\mu, \nu) = \min_{T \in \Pi(\mu,\nu)} \left(\sum_{ij} T_{ij} M_{ij} \right),$$

where the minimum is taken over $T \in \Pi(\mu, \nu)$, i.e., over the space of product distributions of μ and ν. In this case, the value of T that gives the minimum value of the summation is useful to us – it represents how much "probability mass" we need to move from context to context between the pair of words, which tells us why the model ranks the pair of words as it does.

Datasets & Evaluation

Unsupervised data suffices for this system, as it only needs co-occurrence information that can be obtained by sliding a window in one pass through a text corpus. We evaluate the system on three tasks – word similarity, word entailment, and sentence similarity.

The following are the datasets we plan to use:

- Word Similarity – WordSim353
- Word Entailment – HypeNet
- Sentence Similarity – SemEval

Literature Review

This paper describes the system in detail. The system relies on statistical information (co-occurrence counts) to define probability distributions for each word across its contexts (grounded in a learned global embedding space). **Another study** takes a different approach, learning the distributions rather than defining them statically.

Optimal transport theory, while it has been made use of in NLP, is rare in the field of representation learning. Its other uses have been mainly in dataset distances (**Alvarez-Melis & Fusi, 2020**), entity alignment (**Shichao, Lu & Xiangliang, 2019**), and text generation (**Chen et al., 2020**).

Timeline

- Dataset - 12th October, 2022
- Baseline Analysis - 15th October, 2022
- Complete Evaluation Pipeline - 19th October, 2022
- Model Making - 5th November, 2022
- Analysis & Report - 16th November, 2022