A Generative Model of Words and Relationships from Multiple Sources

cBio@MSKCC

@__hylandSL

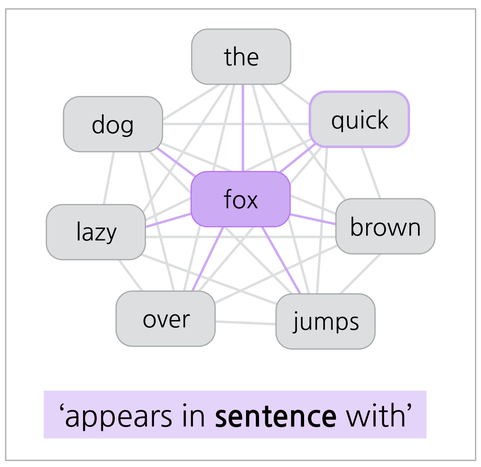
code and data:
https://github.com/corcra/bf2

Stephanie L. Hyland[1,2], Theofanis Karaletsos[1], Gunnar Rätsch[1]
[1]Computational Biology, Memorial Sloan Kettering Cancer Centre, New York
[2]Weill Cornell Graduate School of Medical Sciences, New York

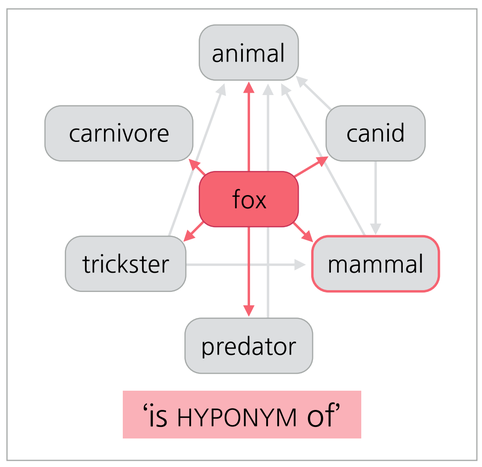
Introduction

- **Word embeddings:** useful semantic representations for downstream natural language processing tasks:
  - distance in embedding space ~ semantic distance
  - learn using **co-occurrence statistics**
- Technical domains (such as **medicine**) may have:
  - x corpora of limited size and expressivity
  - ✓ prior knowledge encoded in knowledge graphs

unstructured text



structured knowledge



Model

- **Approach:**
  1. combine **structured** (knowledge graph) and **unstructured** (free-text) data
  2. generalise co-occurrence to include **relationships** defined by **edges** in graph: represent as **affine transformations**
- Joint generative model of relationships and entity pairs
- Boltzmann distribution:

$$P(S, R, T|\Theta) = \frac{1}{Z(\Theta)} e^{-\mathcal{E}(S,R,T|\Theta)}$$

tamoxifen CAN_TREAT *breast_cancer*
dog APPEARS_IN_SENTENCE_WITH *fox*
brain IS_LOCATION_OF *glioma*
fox IS_HYPONYM_OF *carnivore*

$$\mathcal{E}(S, R, T|\Theta) = -\frac{\mathbf{v}_T \cdot G_R \mathbf{c}_S}{\|\mathbf{v}_T\| \|G_R \mathbf{c}_S\|}$$

(cosine similarity)

"**S** is related to **T** through **R**"

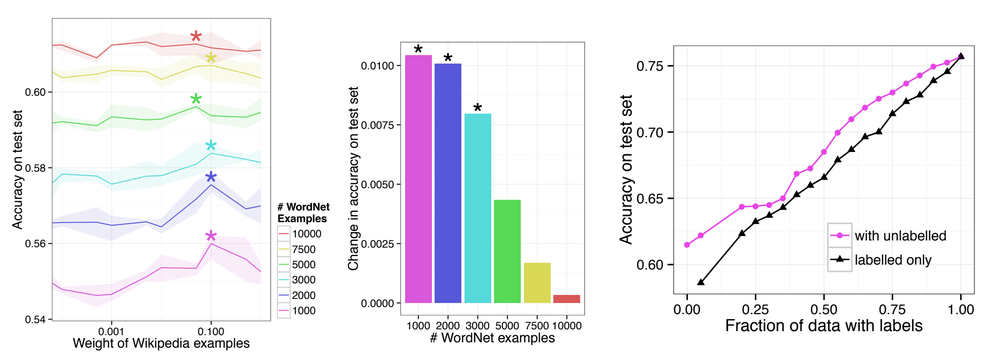
- **Latent** relationships from unlabelled examples
- Inference: **stochastic maximum likelihood** (PCD)

Experiments

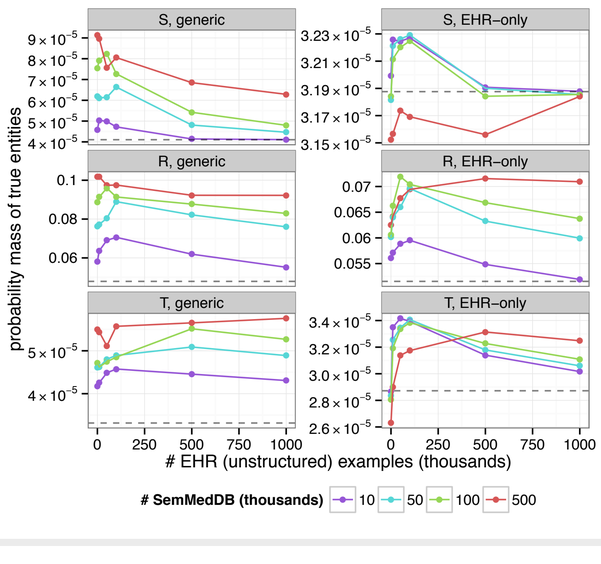
Triplet **classification**: predict if (**S, R, T**) is true/false:

A. unstructured data helps when structured data is scarce

B. **unlabelled** structured data also helps



Edge prediction: predict **R** given **S** and **T**, etc.: report total probability of all correct responses



Knowledge transfer: we can predict relationships between entities appearing **only** in the unstructured (EHR) data

Data

- **Generic English:** Wikipedia (unstructured), WordNet (structured)
  - 12 relationships (including APPEARS_IN_SENTENCE)
  - 112,581 WordNet training examples
- **Medical English**: electronic health records (EHR) from MSKCC (unstructured), SEMMEDDB (structured)
  - took top 20 relationships from SEMMEDDB
  - identified UMLS concepts in EHR

Conclusion

- Our model learns embeddings using both **distributional statistics** and structured **knowledge graphs**
- Relationships between words are **affine transformations** of the space
- Combining data sources can improve the quality of embeddings
- We can predict relationships for entities **not appearing** in the graph

Kilicoglu *et al*. 2012. SemMedDB: a PubMed-scale repository of biomedical semantic predications. (*Bioinformatics)*
Mikolov *et al.* 2013. Distributed representations of words and phrases and their compositionality. (*NIPS)*
Tieleman, T. 2008. Training restricted Boltzmann machines using approximations to the likelihood gradient. *(ICML)*.
Socher *et al*. 2013. Reasoning with neural tensor networks for knowledge base completion. *(NIPS)*

paper: