Biomedical relation inference via embeddings

Alex Derry, Department of Biomedical Informatics
aderry@stanford.edu

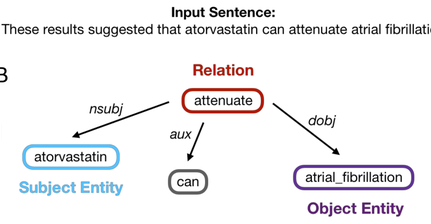
Problem Statement

- Understanding the knowledge graph of relationships between drugs, genes, and diseases is crucial to biomedical discovery
- Scientific knowledge is distributed across millions of articles of unstructured text
- Previous approaches based on co-occurrence, pattern matching, or dependency parsing suffer from low recall and lack of ability to model latent corpus-wide relationship features
- Word embeddings could be an alternative solution that help address such problems

Data

Corpus: ~28.6 m abstracts from PubMed → ~153.3 m sentences
Preprocessing: Phrase concatenation, tokenization, removal of punct.

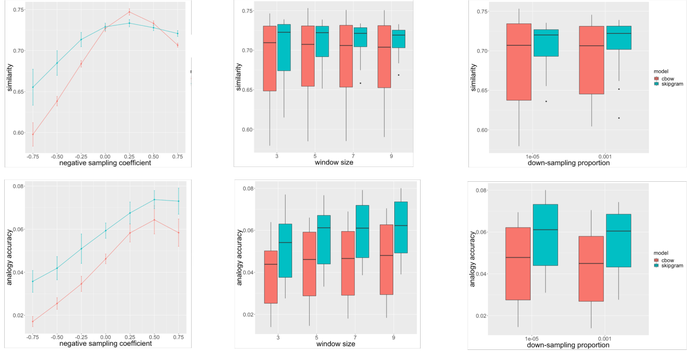
Training data for supervised predictions: Extracted from PharmGKB and Therapeutic Target databases, plus extracted from dependency parses (see right for example)



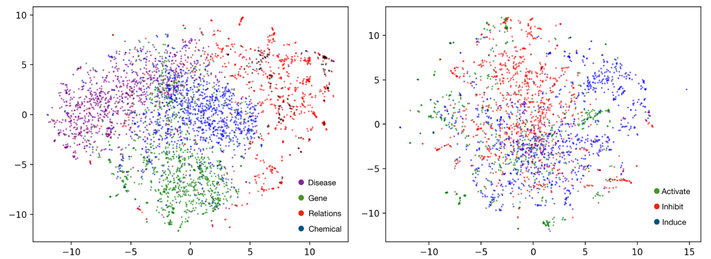
Embedding Model: Word2Vec

Gridsearch: Tuned parameters were negative sampling coefficient, window size, and down-sampling rate.
Constants: Dimensionality (d = 300)
Evaluation: (1) Word pair similarity—Spearman corr. on bio-SimLex dataset; (2) Analogy solving—accuracy on specially generated dataset
Results:

- *Skip-gram* outperforms *CBOW* across the board
- Larger window size is better for analogy modeling
- Negative sampling coefficient of 0.25 or 0.50 is optimal



Embedding Space Evaluation



Results from optimal model after 25 epochs:
tSNE visualizations of the entire embedding space (left) and the relationship space alone (right) show that

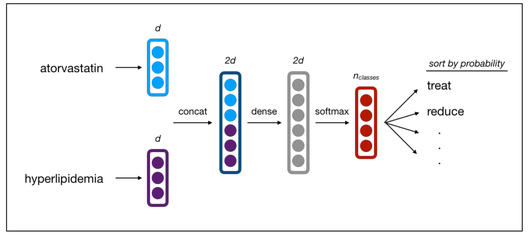
- Diseases, genes, drugs and relations are separable
- Different types of relations are also distinguishable—therefore aggregating within each type allows the model to learn features of each

Note: black dots in the left plot represent passive voice ("inhibited by"), while red dots represent active voice ("inhibits"). These are not separable because Word2Vec cannot model directionality

Prediction Models

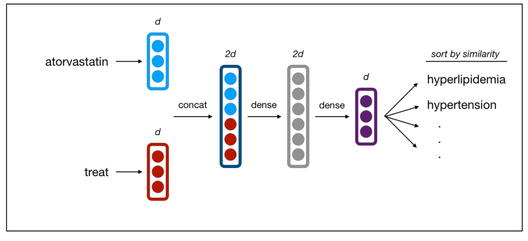
Task 1: Given embeddings of two entities, build a classifier for the relationship between them

- Loss: categorical cross-entropy
- Learning rate: 10^{-5}
- Epochs: 10



Task 2: Given embeddings of a subject entity and a relationship type, predict the object entity

- Loss: cosine distance
- Learning rate: 10^{-5}
- Epochs: 10



Analysis and Results

Evaluation metrics:

1. *Top-n accuracy*: does correct result appear in top *n* predictions?
2. *Mean reciprocal rank:* $MRR = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{1}{r_i}$
3. *Mean average precision:* $MAP = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \frac{1}{AP_i}$ $AP = \frac{1}{n_{true}} \sum_{r=1}^{n_{true}} \max_{r* \geq r} prec(r*)$

	Model 1: Relation classification			**Model 2:** Object Prediction	
	Accuracy (top 1)	**Accuracy (top 10)**	**MRR**	**MAP**	**MRR**
No class aggregation	0.142	0.561	0.267	2.529×10^{-4}	0.004
With class aggregation	**0.231**	**0.784**	**0.393**	**0.042**	**0.273**

Key findings:

1. Averaging representations of equivalent classes significantly improves performance
2. Relation classification was successful—correct relation appeared in top 3 predictions on average
3. Object prediction was much more challenging—decent ability to extract a single correct answer but at the expense of all others

Conclusions

- Word2Vec is able to learn embeddings of genes, diseases, and relationships between them that are useful for representing biomedical interactions, especially after combining the vectors for equivalent relations
- This approach does not rely on co-occurrences and can model novel and/or latent relationships
- Future work includes improving the model for object prediction and incorporating contextual information to model directionality of interactions