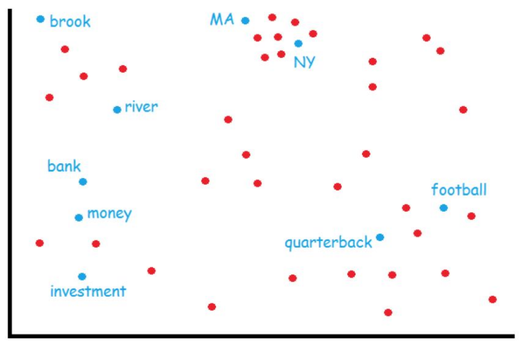
AWE-CM Vectors: Augmenting Word Embeddings with a Clinical Metathesaurus

Willie Boag, Mohamed Kane
MIT CSAIL

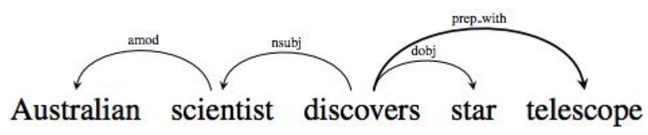
Word Embeddings

word2vec (Mikolov 2013) created massive interest in low-dimensional dense word representations.



The Google News vectors are three orders of magnitude larger than the largest publically-available dataset of clinical notes.
How can we use domain knowledge to augment the smaller clinical text corpora?

Generalized word-context pairs



word2vec Skip-Gram with Negative Sampling estimates how likely a (word,context) pair is to have actually occured in the training data.

$$P(\ (discover, star)\ is\ real\ ) = \sigma(W_{discover} \cdot C_{star})$$

word2vecf (Levy and Goldberg 2014) generalized word2vec to include "contexts" other than adjacent words, but instead dependency parse arcs.

$$P(\ (discover, \textbf{prep\_with:telescope})\ is\ real\ ) = \sigma(W_{discover} \cdot C_{\textbf{prep\_with:telescope}})$$

Approach

Domain knowledge feature extraction as "contexts".
UMLS CUIs as contexts:

$$P(\ (blood, \textbf{C0005767})\ is\ real\ ) = \sigma(W_{blood} \cdot C_{\textbf{C0005767}})$$

Train word2vecf vectors with both adjacent words *and* CUIs as contexts.

2,683,398,577 word-based (w,c) context pairs and
265,699,787 ontology-based (w,CUI) context pairs.

https://github.com/wboag/awecm

Dataset

MIMIC-III v1.4 contains de-identified EHR data from over 58,000 Beth Israel admissions for nearly 38,600 adult patients. The data consists of 2 million notes totaling 500 million tokens.

Each note was preprocessed by:

- removing PHI tags
- collapsing all-caps phrases into a single token
- reducing common age regular expressions to per-decade age tokens
- removing all non-alphanumeric characters
- normalizing all non-age numbers to zero.

Evaluation

MayoSRS. clinician judgments, created by the University of Minnesota

1.0 "sodium" "mri"
1.0 "hand splint" "splinter hemorrhage"
2.15 "immunization" "immunodeficient"
4.38 "swallowing" "peristalsis"
6.85 "ileitis" "Crohns Disease"
8.23 "metastasis" "carcinomatosis"

Embeddings are evaluated by the pearson correlation between clician judgment score and the cosine of the phrase centroids.

Future Work

1. Extract additional features.
   a. tree path relationships from UMLS ontology
   b. character ngrams
2. Demonstrate effect of augmented contexts as a function of corpus size.
3. Evaluate on multiple tasks, including extrinsic (clinical NER), intrinsic (clinical analogies), and qualitative.

Results

Table 1: Spearman coefficient of correlation with various experts.

	MayoSRS	MiniMaySRS: doctors	MiniMaySRS: coders
Google News	0.128	0.145	0.302
MIMIC w2v	**0.398**	0.442	**0.572**
MIMIC W2VF: words	0.324	0.495	0.489
AWE-CM	0.365	**0.508**	0.514

Augmented domain knowledge vectors outperform the text-only word2vecf vectors across the board.

Rows 2 and 3 should, in theory, be the same model. The differences between them indicate differences introduced by implementations into word2vec vs word2vecf.

Acknowledgements

This research was funded in part by the National Science Foundation Graduate Research Fellowship Program grant No. 1122374.