# Machine Learning for Textual and Unstructured Data

## Lecture 3: Word Embeddings

Stephen Hansen
University College London

FUNDACIÓN RAMÓN ARECES

IESE Business School University of Navarra

Center for International Finance

# Introduction

The dimensionality reduction methods in the previous slides rely on global co-occurrence patterns.

In natural language, local context is arguably a more natural guide to meaning: "You shall know a word by the company it keeps" (Firth).

Suppose we wish to understand the meaning of the 'alien word' $%& which has been inserted in place of an English word in the following sentence:

> *Every morning last summer in Greece, I visited the $%& where I would swim, play in the sand, and sunbathe.*

The words present in this text snippet give strong clues about the meaning of $%&.

Words present many sentences away would generally be less informative.

# Word Embeddings

A word embedding is a low-dimensional vector representation of a word.

Ideally in this low-dimensional vector space words with similar meanings will lie close together.

The construction of word embeddings has been a major topic in NLP in the past decade.

Embedding vectors can be of interest in their own right, or else form part of the representation of a document for other tasks.

# Local Contexts and Word Embeddings

Recall that $w_{d,n}$ is the $n$th word in document $d$.

The *context* of $w_{d,n}$ is a length-$2L$ window of words around $w_{d,n}$:

$$C(w_{d,n}) = [w_{d,n-L}, w_{d,n-L+1}, \dots, w_{d,n+L-1}, w_{d,n+L}]$$

Can truncate context appropriately if window stretches past beginning or end of text.

In line with Firth's distributional hypothesis, modern word embedding models seek to generate similar embeddings for words that share similar contexts.

# GloVe

The GloVe model [Pennington et al., 2014] begins with a $V \times V$ matrix $\mathbf{W}$ of local word co-occurrences.

$W_{ij}$ is the number of times term $j$ appears within the context of $i$.

Assign to each term $v$ an embedding vector $\boldsymbol{\rho}_v \in \mathbb{R}^V$.

$$\min \sum_{i,j} f(W_{i,j}) \left(\boldsymbol{\rho}_i^T \boldsymbol{\rho}_j - \log\left(W_{i,j}\right)\right)^2$$

Terms that co-occur frequently will have more highly correlated embedding vectors.

# Word2Vec

Word2vec [Mikolov et al., 2013a, Mikolov et al., 2013b] is another well-known model for word embeddings that incorporates local context.

In addition to an embedding vector, each term is assigned a context vector $\boldsymbol{\alpha}_v \in \mathbb{R}^V$.

Word vectors are chosen to solve word-prediction tasks:

$$\Pr\left[\, w_{d,n} = v \mid C(w_{d,n}) \,\right] = \frac{\exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_v)}{\sum_{v'} \exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_{v'})} \text{ where } \overline{\boldsymbol{\alpha}}_{d,n} = \frac{1}{2L} \sum_{w \in C(w_{d,n})} \boldsymbol{\alpha}_w$$

Example of self-supervised learning.

The version of Word2Vec is the continuous-bag-of-words model; the skip-gram model instead predicts context words given a center word.

# Terms Close to Uncertainty in FOMC Transcripts

| term | sim |
|---|---|
| uncertainties | 0.741 |
| anxiety | 0.48 |
| pessimism | 0.479 |
| skepticism | 0.465 |
| optimism | 0.445 |
| caution | 0.442 |
| gloom | 0.437 |
| uncertain | 0.433 |
| sensitivity | 0.427 |
| angst | 0.426 |

| term | sim |
|---|---|
| challenges | 0.415 |
| fragility | 0.405 |
| clarity | 0.401 |
| concerns | 0.4 |
| risks | 0.397 |
| disagreement | 0.387 |
| volatility | 0.384 |
| tension | 0.383 |
| certainty | 0.382 |
| skepticism | 0.38 |

# Terms Close to Risk

| term | sim |
|---|---|
| risks | 0.737 |
| threat | 0.609 |
| danger | 0.541 |
| dangers | 0.463 |
| vulnerability | 0.457 |
| chances | 0.451 |
| breakout | 0.433 |
| probability | 0.426 |
| possibility | 0.409 |
| likelihood | 0.406 |

| term | sim |
|---|---|
| misdirected | 0.385 |
| odds | 0.379 |
| uncertainty | 0.375 |
| concern | 0.371 |
| prospect | 0.37 |
| instability | 0.363 |
| potentially | 0.352 |
| concerns | 0.352 |
| challenges | 0.346 |
| risking | 0.342 |

# References I


Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient Estimation of Word Representations in Vector Space.
arXiv:1301.3781 [cs].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed Representations of Words and Phrases and their Compositionality.
arXiv:1310.4546 [cs, stat].

Pennington, J., Socher, R., and Manning, C. (2014).
GloVe: Global Vectors for Word Representation.
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.