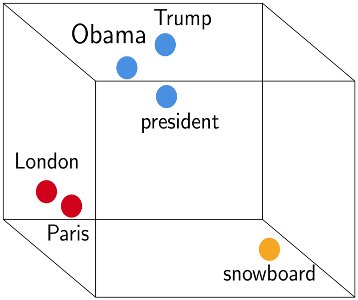
Word Embeddings for Entity-annotated Texts

Satya Almasian, Andreas Spitz and Michael Gertz

April 16, 2019

Heidelberg University
Institute of Computer Science
Database Systems Research Group
almasian@stud.uni-heidelberg.de

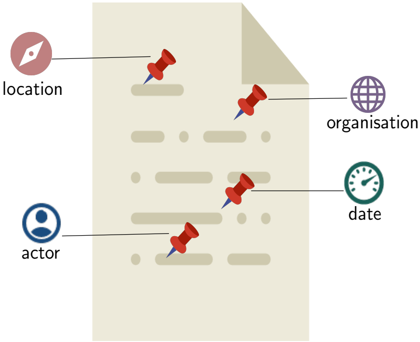
Motivation



Word Embeddings:

- Word represented as vectors of real numbers
- Words with similar meaning $\Rightarrow$ mapped to the nearby points

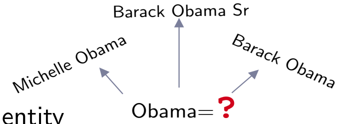
What about Named Entities?



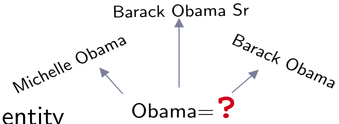
- Named entities: Words that belong to a set of pre-defined classes (person, location, organisation . . .).
- Normal word embeddings
  $\Rightarrow$ treat all the words equally

Challenges of Normal Word Embeddings for Named Entities

- Fails to distinguish between multiple **homographs** of an entity

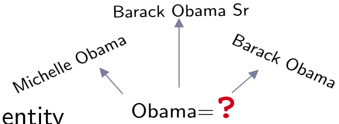


Challenges of Normal Word Embeddings for Named Entities



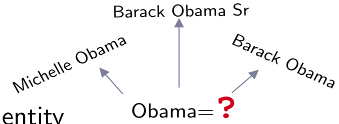
- Fails to distinguish between multiple **homographs** of an entity
- Cannot recognize **compound words** $\Rightarrow$ $\boxed{\text{Barack}}$ $\boxed{\text{Obama}}$

Challenges of Normal Word Embeddings for Named Entities

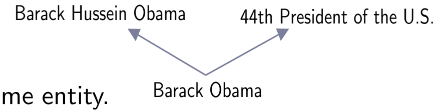


- Fails to distinguish between multiple **homographs** of an entity
- Cannot recognize **compound words** ⇒ $\boxed{\text{Barack}}$ $\boxed{\text{Obama}}$
- Inability to deal with date information, e.g., 1 January 1998 or 1998-01-01

Challenges of Normal Word Embeddings for Named Entities



- Fails to distinguish between multiple **homographs** of an entity
- Cannot recognize **compound words** ⇒ Barack | Obama
- Inability to deal with date information, e.g., 1 January 1998 or 1998-01-01



- Unable to detect multiple **surface forms** of the same entity.

Entity Embeddings

Entity and Word Embeddings

Annotate given text with named entities

Entity and Word Embeddings

Annotate given text with named entities

Training state-of-the-art word embedding
(word2vec, GloVe) on annotated text

Entity and Word Embeddings

Annotate given text with named entities

Training state-of-the-art word embedding (word2vec, GloVe) on annotated text

1. Extract word co-occurrence graph from text
2. Embed the nodes of the co-occurrence graph with node embedding models (DeepWalk, VERSE)

Word Embeddings of Annotated Text

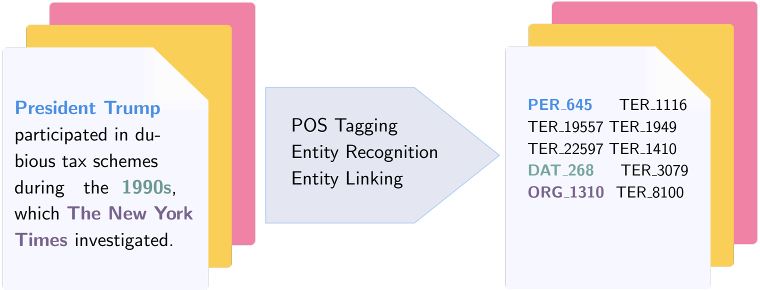
President Trump participated in dubious tax schemes during the **1990s**, which **The New York Times** investigated.

Word Embeddings of Annotated Text

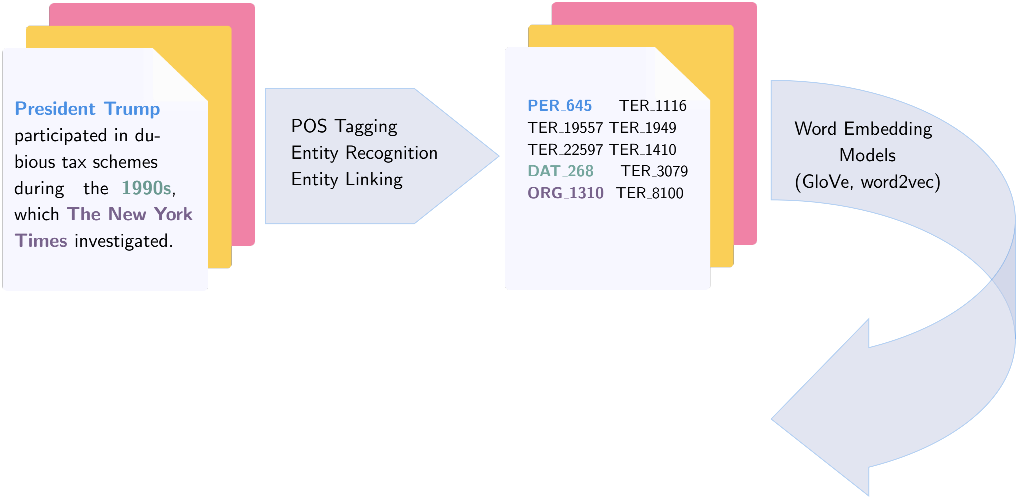
President Trump participated in dubious tax schemes during the **1990s**, which **The New York Times** investigated.

POS Tagging
Entity Recognition
Entity Linking

Word Embeddings of Annotated Text



Word Embeddings of Annotated Text



Word Embeddings of Annotated Text

President Trump participated in dubious tax schemes during the **1990s**, which **The New York Times** investigated.

POS Tagging
Entity Recognition
Entity Linking

PER_645 TER_1116
TER_19557 TER_1949
TER_22597 TER_1410
DAT_268 TER_3079
ORG_1310 TER_8100

Word Embedding Models (GloVe, word2vec)

	Dimensions					
PER_645	-0.1	-0.31	0.13	0.7	1.3	$\cdots$
TER_1949	0.1	1.23	8.7	4.7	0.2	$\cdots$
DAT_268	-0.41	-2.31	-0.6	0.7	-0.1	$\cdots$
ORG_1310	2.1	4.31	0.53	-0.7	0.9	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Motivation for Co-occurrence Graphs

President Trump participated in dubious tax schemes during the 1990s, including instances of outright fraud, that greatly increased the fortune he received from his parents, an investigation by The New York Times has found. He won the presidency proclaiming himself a self-made billionaire, and he has long insisted that his father, the legendary New York City builder **Fred C. Trump**, provided almost no financial help.

Sentences: 2
Words: 53

- **Entity-entity relations** are captured better using sentence-based distances.

Motivation for Co-occurrence Graphs

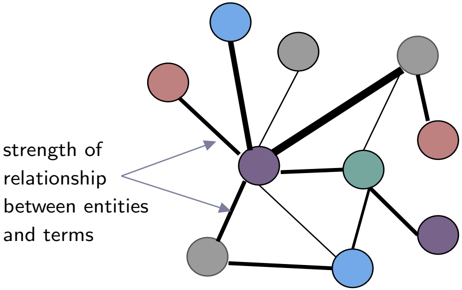
President Trump participated in dubious tax schemes during the 1990s, including instances of outright fraud, that greatly increased the fortune he received from his parents, an investigation by The New York Times has found. He won the presidency proclaiming himself a self-made billionaire, and he has long insisted that his father, the legendary New York City builder **Fred C. Trump**, provided almost no financial help.

Sentences: 2

Words: 53

- **Entity-entity relations** are captured better using sentence-based distances.
- **Terms** are less likely to be related to entities outside of their own sentence.

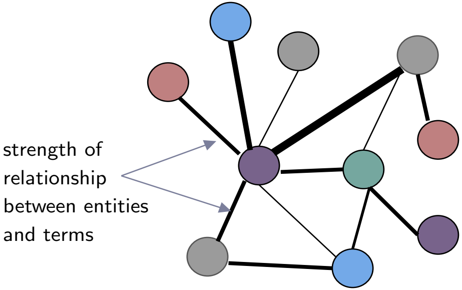
Weighted Entity Co-occurrence Graph



- Represents a document collection as an entity co-occurrence graph [1]

[1] Spitz, A., Gertz, M.: Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. SIGIR 2016.

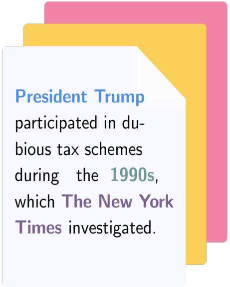
Weighted Entity Co-occurrence Graph



- Represents a document collection as an entity co-occurrence graph [1]
- **Entity-entity** relations ⇒ **sentence-based** weighting function
- **Entity-term** relations ⇒ **word-based** weighting function

[1] Spitz, A., Gertz, M.: Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. SIGIR 2016.

Node Embeddings of Co-occurrence Graphs



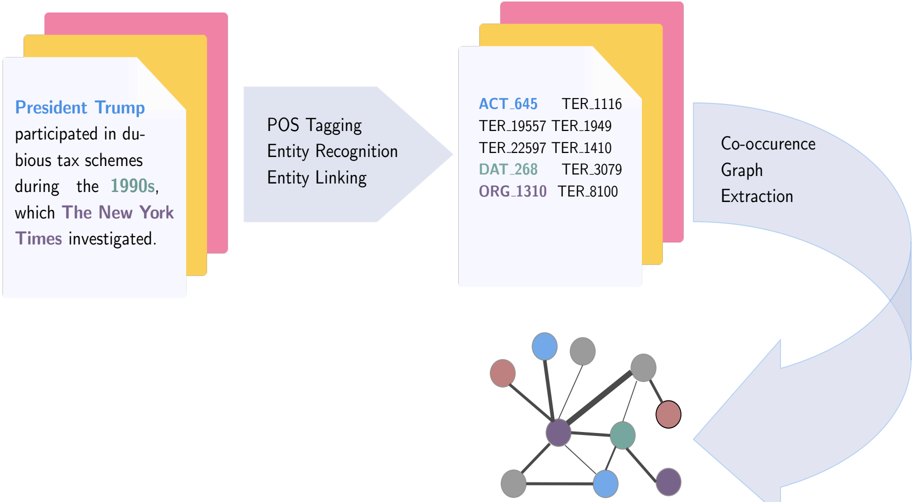
Node Embeddings of Co-occurrence Graphs

President Trump participated in dubious tax schemes during the 1990s, which The New York Times investigated.

POS Tagging
Entity Recognition
Entity Linking

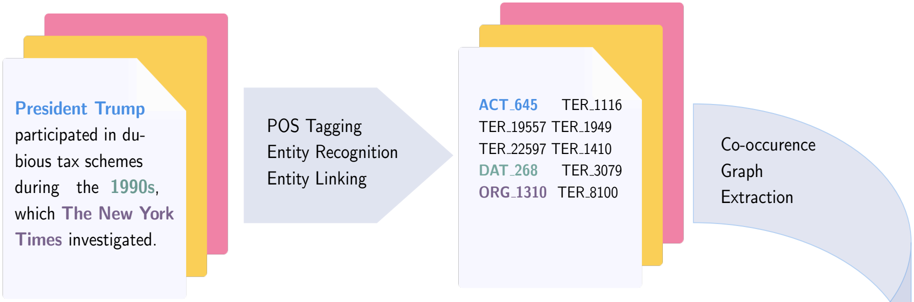
ACT_645 TER_1116
TER_19557 TER_1949
TER_22597 TER_1410
DAT_268 TER_3079
ORG_1310 TER_8100

Node Embeddings of Co-occurrence Graphs

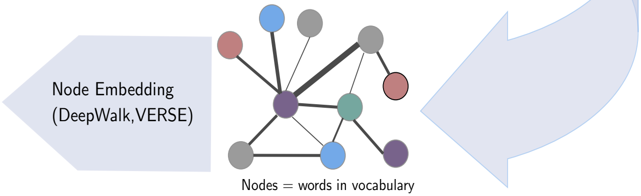


Nodes = words in vocabulary

Node Embeddings of Co-occurrence Graphs

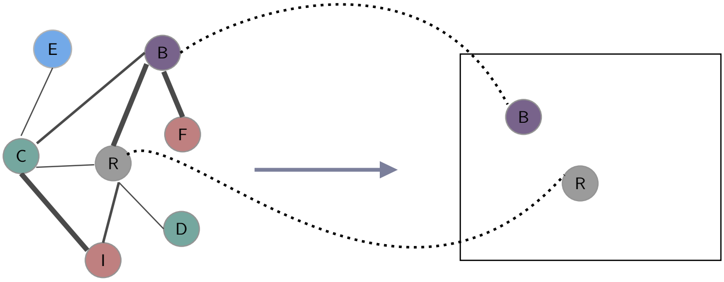


	Dimensions				
ACT_645	-0.1	-0.31	0.13	0.7	$\cdots$
TER_1949	0.1	1.23	8.7	4.7	$\cdots$
DAT_268	-0.41	-2.31	-0.6	0.7	$\cdots$
ORG_1310	2.1	4.31	0.53	-0.7	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	



Node Embeddings

- Encode nodes so that similarity in the embedding space approximates similarity in the original network.

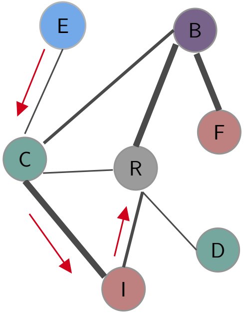


Original Graph

Embedding space

Node Embeddings

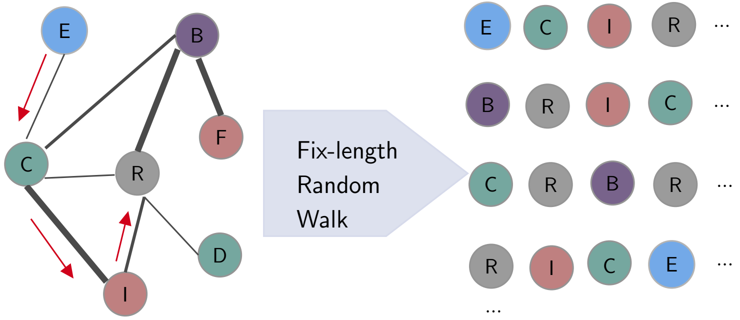
DeepWalk (Perozzi et al.):



Fix-length Random Walk

Transition probability proportional to the edge weights

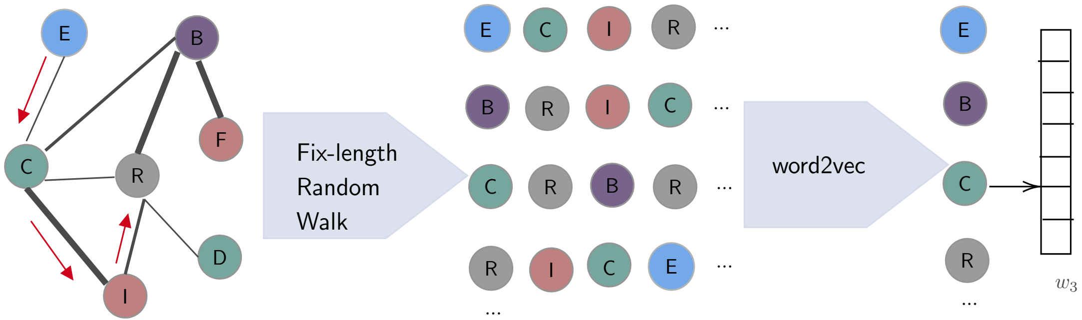
DeepWalk (Perozzi et al.):



Transition probability proportional to the edge weights

Corpus:
Nodes ⇒ words
Random walks ⇒ sentences

DeepWalk (Perozzi et al.):



Transition probability proportional to the edge weights

Corpus:
Nodes ⇒ words
Random walks ⇒ sentences

Embeddings

VERSE (Tsitsulin et al.):

- Learns embeddings by training a **single-layer neural network**

Node Embeddings

VERSE (Tsitsulin et al.):

- Learns embeddings by training a **single-layer neural network**
- Accepts **diverse** similarity measures

Node Embeddings

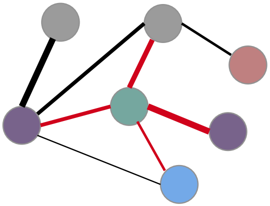
VERSE (Tsitsulin et al.):

- Learns embeddings by training a **single-layer neural network**
- Accepts **diverse** similarity measures
- Minimizes **KL-divergence** from the given similarity distribution sim_G to similarity in embedding space sim_E
  - $\sum_{v \in V} KL(sim_G(v,.), sim_E(v,.))$

Node Embeddings

VERSE (Tsitsulin et al.):

- Learns embeddings by training a **single-layer neural network**
- Accepts **diverse** similarity measures
- Minimizes **KL-divergence** from the given similarity distribution sim_G to similarity in embedding space sim_E
  - $\sum_{v \in V} KL(sim_G(v, .), sim_E(v, .))$
- We choose **Adjacency Similarity** as our similarity measure



Only the immediate neighbors are taken into account

Evaluation

Evaluation of Word Embeddings

Relatedness or Similarity:

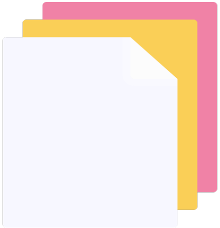
- Datasets containing scores for:
  $\Rightarrow$ **relatedness**: association of words
  $\Rightarrow$ **similarity**: degree of synonymity
- The cosine similarity should have a high correlation with human scores.

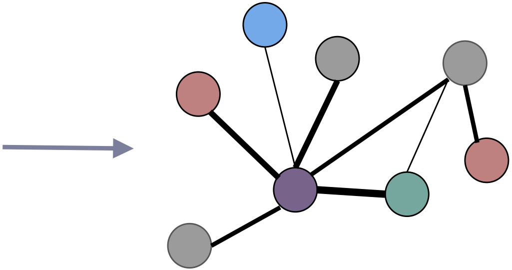
Analogy:

- Datasets containing:
  $\Rightarrow$ word pairs with similar relations
- Task: Given (a, b, x), find y where, $x : y$ resembles $a : b$
- **Berlin** is to **Germany**, as **London** is to ?
  $\rightarrow$ **England**

Data

Training:





200K english news articles,
from June to November 2016

Co-occurence graph,
93K nodes,
9 million edges

Testing:

4 Similiarity datasets
3 Relatedness datasets
2 Analogy datasets

Relatedness or Similarity

Correlation with human judgements:

	raw text		annotated text			
Datasets	word2vec	GloVe	word2vec	GloVe	DeepWalk	VERSE
Word Similarity	**0.551**	0.410	0.550	0.394	0.434	0.492
Word Relatedness	0.491	0.378	0.490	0.380	0.432	**0.510**
average	**0.526**	0.400	0.524	0.389	0.433	0.500

- Performance of word embedding models **degrades** after annotation.

Relatedness or Similarity

Correlation with human judgements:

Datasets	raw text: word2vec	raw text: GloVe	annotated text: word2vec	annotated text: GloVe	annotated text: DeepWalk	annotated text: VERSE
Word Similarity	**0.551**	0.410	0.550	0.394	0.434	0.492
Word Relatedness	0.491	0.378	0.490	0.380	0.432	**0.510**
average	**0.526**	0.400	0.524	0.389	0.433	0.500

- Performance of word embedding models **degrades** after annotation.
- **Graph-based embeddings (VERSE)** $\Rightarrow$ better for **relatedness** tasks
- **Normal word embeddings (word2vec)** $\Rightarrow$ better for **similarity** tasks.

Analogy

Accuracy of prediction:

	raw text		annotated text			
Datasets	word2vec	GloVe	word2vec	GloVe	DeepWalk	VERSE
Google Analogy	0.013	0.019	0.003	0.015	0.009	**0.035**
Microsoft Research	0.014	**0.019**	0.001	0.014	0.002	0.012
average	0.013	0.019	0.002	0.014	0.005	**0.023**

- **Graph-based embedding (VERSE)** $\Rightarrow$ better for datasets containing named entities
- **Normal word embedding (GloVe)** $\Rightarrow$ better for term-based datasets

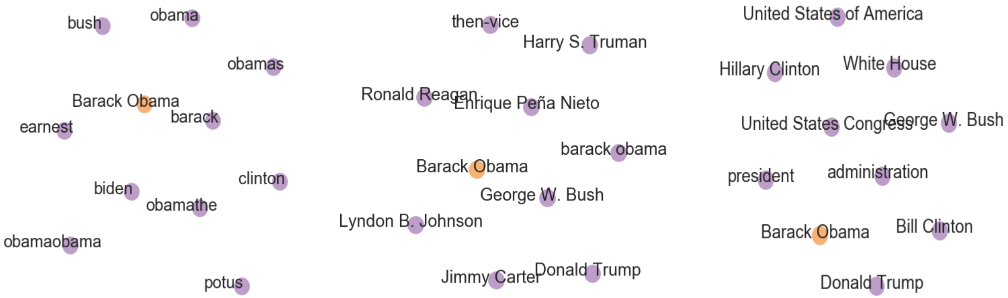
Analogy

Accuracy of prediction:

	raw text		annotated text			
Datasets	word2vec	GloVe	word2vec	GloVe	DeepWalk	VERSE
Google Analogy	0.013	0.019	0.003	0.015	0.009	**0.035**
Microsoft Research	0.014	**0.019**	0.001	0.014	0.002	0.012
average	0.013	0.019	0.002	0.014	0.005	**0.023**

- **Graph-based embedding (VERSE)** $\Rightarrow$ better for datasets containing named entities
- **Normal word embedding (GloVe)** $\Rightarrow$ better for term-based datasets
- **Typed search:** Only look at candidates that share the same type as entities in question
  - **Entity-centric (location)** questions: VERSE predicts $1,662$ (24.1%)
  - word2vec $\Rightarrow$ 14 (0.20%) and GloVe $\Rightarrow$ 16 (0.23%).

Neighbourhood Structure for “Barack Obama”



(a) word2vec(raw) **(b)** word2vec(annotated) **(c)** VERSE

Conclusion

Conclusion: Usability of Embeddings

When to use which embedding for which task?

Normal Word Embeddings | Graph-based Embeddings

Conclusion: Usability of Embeddings

When to use which embedding for which task?

Find similar and synonymous words

Normal Word Embeddings

Analogy between terms

Graph-based Embeddings

Conclusion: Usability of Embeddings

When to use which embedding for which task?

Find similar and synonymous words

Normal Word Embeddings

Analogy between terms

Find related and associated words

Graph-based Embeddings

Analogy between entities

Analysis of entity entity relations

Outlook

- Use embeddings as **input features** for entity-centric tasks that benefit from relatedness relations
  - $\Rightarrow$ Query expansion
  - $\Rightarrow$ Name entity recognition and linkage

Outlook

- Use embeddings as **input features** for entity-centric tasks that benefit from relatedness relations
  - $\Rightarrow$ Query expansion
  - $\Rightarrow$ Name entity recognition and linkage
- Creating datasets containing named entities for evaluation tasks

Thank You

The code:

https://github.com/satya77/Entity_Embedding