Intrinsic Subspace Evaluation of Word Embedding Representations

Yadollah Yaghoobzadeh, Hinrich Schütze

ACL 2016

CIS - LMU Munich

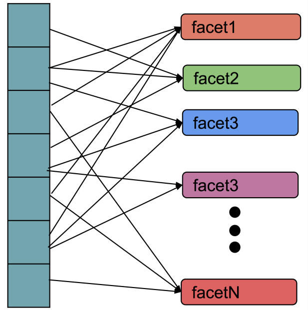
Outline

1. Background
2. Intrinsic Subspace Evaluation
3. Extrinsic Evaluation: Entity Typing

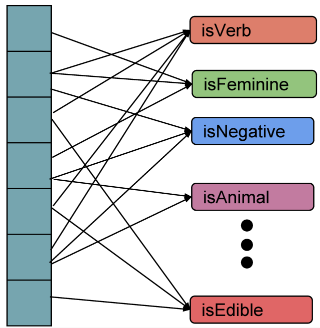
Background

Word Embeddings

... represent generic word properties (facets) in real valued vectors

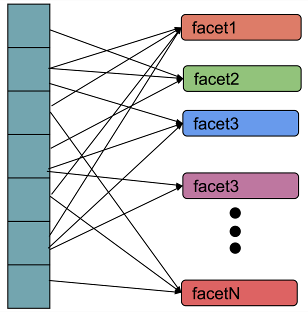


Word Embeddings: Example Facets

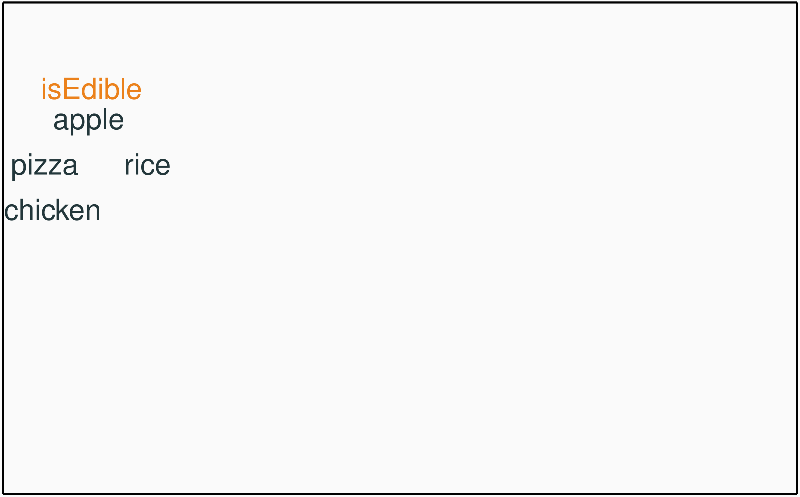


Fullspace vs. Subspace Similarity of Words

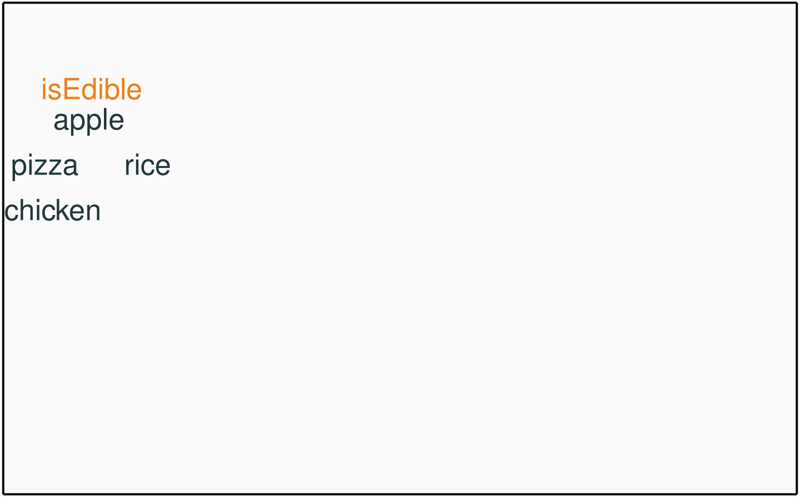
- Fullspace similarity: similarity on all facets
- Subspace similarity: similarity on a subset of facets



What embeddings should ideally represent

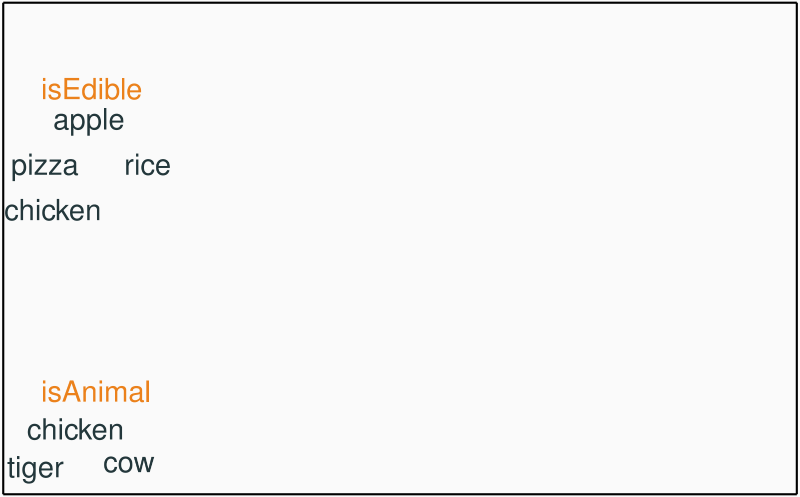


What embeddings should ideally represent



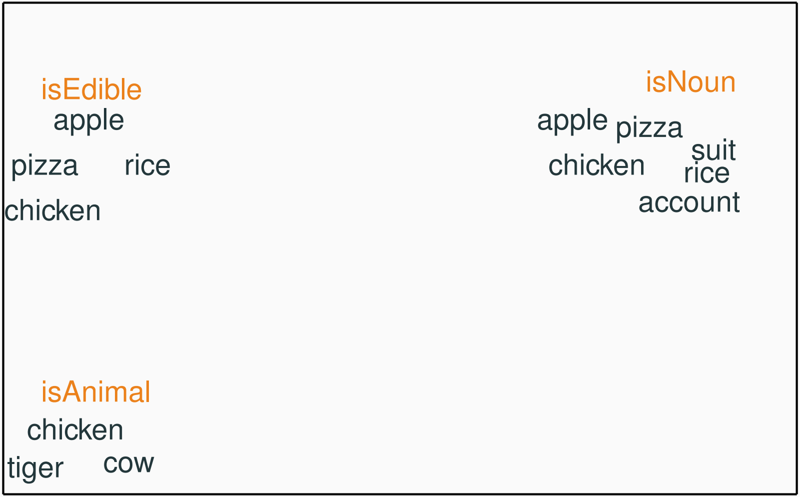
- Ideally, words with similar facets should be close.

What embeddings should ideally represent



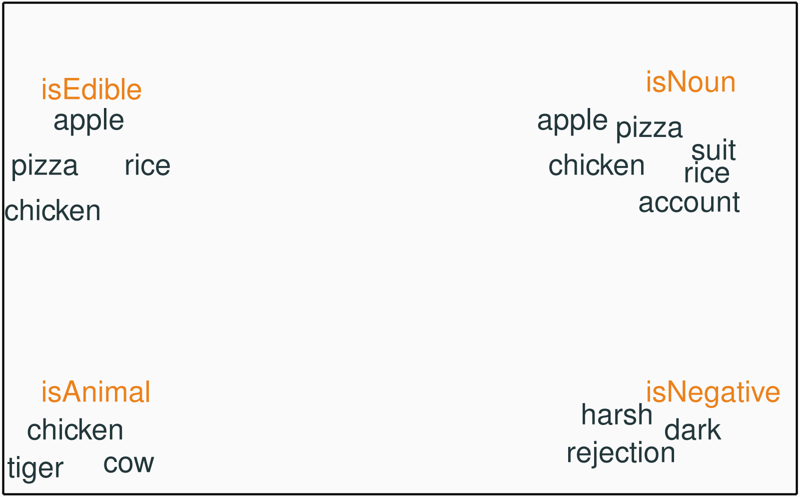
- Ideally, words with similar facets should be close.

What embeddings should ideally represent



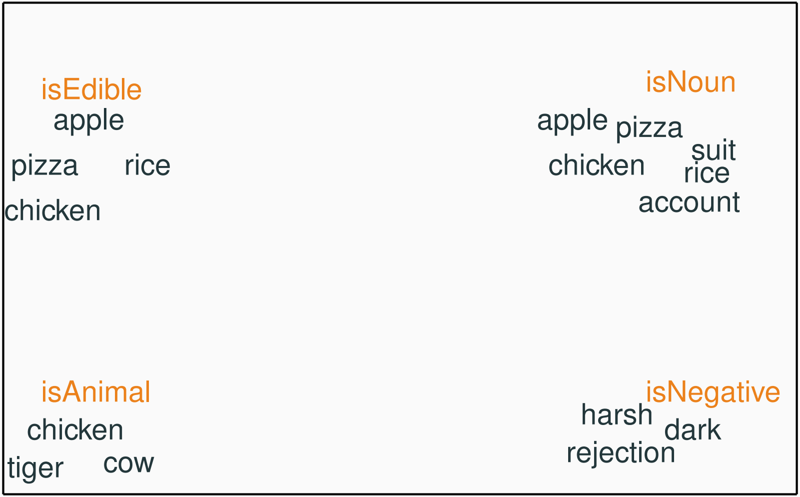
- Ideally, words with similar facets should be close.

What embeddings should ideally represent



- Ideally, words with similar facets should be close.

What embeddings should ideally represent



- Ideally, words with similar facets should be close.
- But they are not on fullspace sim, only on subspace sim!

Evaluation of Word Embeddings

Evaluation of Word Embeddings

- Intrinsic:

Evaluation of Word Embeddings

- Intrinsic:
    - Evaluating embeddings generically

Evaluation of Word Embeddings

- Intrinsic:
    - Evaluating embeddings generically
    - Common approach: check fullspace similarities of words

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems
- Extrinsic:

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems
- Extrinsic:
  - Evaluate facets in different NLP tasks

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems
- Extrinsic:
  - Evaluate facets in different NLP tasks
  - Mostly supervised: subspace evaluation

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems
- Extrinsic:
  - Evaluate facets in different NLP tasks
  - Mostly supervised: subspace evaluation
  - Expensive evaluation

Evaluation of Word Embeddings

- Intrinsic:
  - Evaluating embeddings generically
  - Common approach: check fullspace similarities of words
  - Unsupervised: no good way to explore subspaces
  - Hard to analyze
  - Many other problems
- Extrinsic:
  - Evaluate facets in different NLP tasks
  - Mostly supervised: subspace evaluation
  - Expensive evaluation
  - Hard to analyze the results

What Should We Do Then?

We need an intrinsic evaluation that explores subspaces!

Intrinsic Subspace Evaluation

Word Embedding Challenges?!

Word Embedding Challenges?!

- Ambiguity

Word Embedding Challenges?!

- Ambiguity
- Multifacetedness

Word Embedding Challenges?!

- Ambiguity
- Multifacetedness
- Sparsity

Word Embedding Challenges?!

- Ambiguity
- Multifacetedness
- Sparsity
- Conflation

Word Embedding Challenges?!

- Ambiguity
- Multifacetedness
- Sparsity
- Conflation
- ...

Word Embedding Challenges?!

- Ambiguity
- Multifacetedness
- Sparsity
- Conflation
- ...

We can evaluate word embeddings with respect to these challenges!

New Evaluation Methodology (1)

New Evaluation Methodology (1)

- Model the embedding challenges using PCFG grammars

New Evaluation Methodology (1)

- Model the embedding challenges using PCFG grammars
- Perform supervised classification that needs to find proper subspaces to cover the challenges!

New Evaluation Methodology (1)

- Model the embedding challenges using PCFG grammars
- Perform supervised classification that needs to find proper subspaces to cover the challenges!
- Use Corpus-based supervision

New Evaluation Methodology (2)

New Evaluation Methodology (2)

- Define PCFG grammar

New Evaluation Methodology (2)

- Define PCFG grammar
- PCFG models challenges (e.g., ambiguity) in natural language

New Evaluation Methodology (2)

- Define PCFG grammar
- PCFG models challenges (e.g., ambiguity) in natural language
- Generate a corpus using the PCFG

New Evaluation Methodology (2)

- Define PCFG grammar
- PCFG models challenges (e.g., ambiguity) in natural language
- Generate a corpus using the PCFG
- Train embedding models on corpus

New Evaluation Methodology (2)

- Define PCFG grammar
- PCFG models challenges (e.g., ambiguity) in natural language
- Generate a corpus using the PCFG
- Train embedding models on corpus
- Evaluate embeddings using a supervised task

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Challenge 1: Ambiguity

$P(AV_1B \mid S)$	=	$9/20$	
$P(CW_1D \mid S)$	=	$9/20$	
$P(AW_2B \mid S)$	=	$(1-\beta) \cdot 1/20$	
$P(CW_2D \mid S)$	=	$\beta \cdot 1/20$	
$P(a_i \mid A)$	=	$1/10$	$0 \leq i \leq 9$
$P(b_i \mid B)$	=	$1/10$	$0 \leq i \leq 9$
$P(c_i \mid C)$	=	$1/10$	$0 \leq i \leq 9$
$P(d_i \mid D)$	=	$1/10$	$0 \leq i \leq 9$
$P(v_i \mid V_1)$	=	$1/45$	$0 \leq i \leq 49$
$P(w_i \mid W_1)$	=	$1/45$	$5 \leq i \leq 49$
$P(w_i \mid W_2)$	=	$1/5$	$0 \leq i \leq 4$

Two types of contexts: A-B contexts and C-D contexts.

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Two types of contexts: A-B contexts and C-D contexts.

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Two types of contexts: A-B contexts and C-D contexts.

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Unambiguous: $v_0 \ldots v_{49}$ only occur in A-B contexts.

Challenge 1: Ambiguity

$P(AV_1B\|S)$	=	9/20	
$P(CW_1D\|S)$	=	9/20	
$P(AW_2B\|S)$	=	$(1-\beta)\cdot 1/20$	
$P(CW_2D\|S)$	=	$\beta\cdot 1/20$	
$P(a_i\|A)$	=	1/10	$0 \leq i \leq 9$
$P(b_i\|B)$	=	1/10	$0 \leq i \leq 9$
$P(c_i\|C)$	=	1/10	$0 \leq i \leq 9$
$P(d_i\|D)$	=	1/10	$0 \leq i \leq 9$
$P(v_i\|V_1)$	=	1/45	$0 \leq i \leq 49$
$P(w_i\|W_1)$	=	1/45	$5 \leq i \leq 49$
$P(w_i\|W_2)$	=	1/5	$0 \leq i \leq 4$

Unambiguous: $w_5 \ldots w_{49}$ only occur in C-D contexts.

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Ambiguous: $w_0 \dots w_4$ occur in A-B and C-D contexts.

Challenge 1: Ambiguity

$$
\begin{array}{lcll}
P(AV_1B|S) & = & 9/20 & \\
\hline\hline
P(CW_1D|S) & = & 9/20 & \\
P(AW_2B|S) & = & (1-\beta)\cdot 1/20 & \\
P(CW_2D|S) & = & \beta\cdot 1/20 & \\
\hline\hline
P(a_i|A) & = & 1/10 & 0 \leq i \leq 9 \\
P(b_i|B) & = & 1/10 & 0 \leq i \leq 9 \\
P(c_i|C) & = & 1/10 & 0 \leq i \leq 9 \\
P(d_i|D) & = & 1/10 & 0 \leq i \leq 9 \\
\hline
P(v_i|V_1) & = & 1/45 & 0 \leq i \leq 49 \\
P(w_i|W_1) & = & 1/45 & 5 \leq i \leq 49 \\
P(w_i|W_2) & = & 1/5 & 0 \leq i \leq 4
\end{array}
$$

Ambiguity level controlled by β

Ambiguity: Experiment

Ambiguity: Experiment

- SVM classification:
  “can this word occur in a C-D context?”

Ambiguity: Experiment

- SVM classification:
  “can this word occur in a C-D context?”
- The test set: ambiguous words

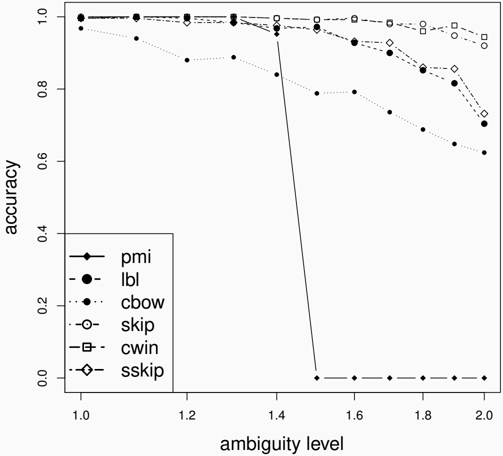
Ambiguity: Experiment

- SVM classification:
  “can this word occur in a C-D context?”
- The test set: ambiguous words
- The train set: all other words

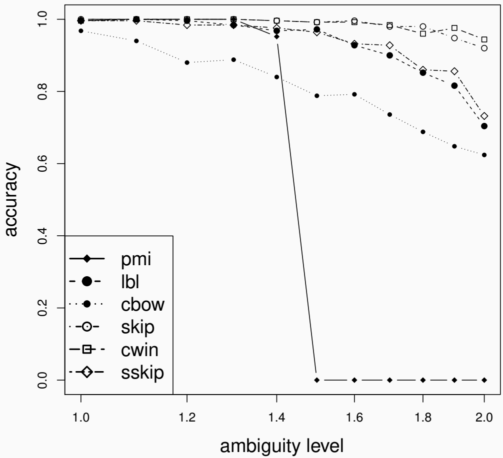
Ambiguity: Experiment

- SVM classification:
  “can this word occur in a C-D context?”
- The test set: ambiguous words
- The train set: all other words
- 50 trials of this experiment for different ambiguity levels

Ambiguity: Comparison of Six Embedding Models



Ambiguity: Comparison of Six Embedding Models



- Accuracy
  - ∼100% for ambiguity level 1.0 (two senses equal)
  - much lower for ambiguity level 2.0 (one sense three times more frequent)

Discussion: Common Premise for Ambiguity

We need k vectors for a word with k senses

Discussion: Common Premise for Ambiguity

We need k vectors for a word with k senses

Why?

Why Some People Are Against single-embedding-per-word

- Two senses of “suit”: litigation vs. clothing

Why Some People Are Against single-embedding-per-word

outfit

$\vec{s}_2$ clothing

legal-case

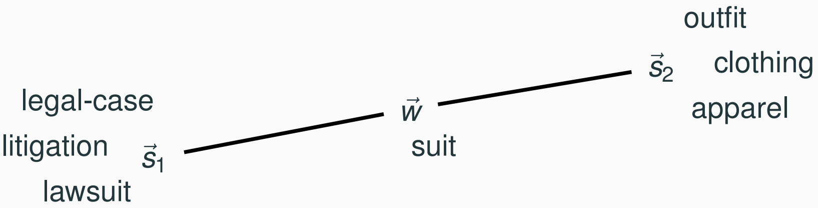
apparel

litigation $\vec{s}_1$

lawsuit

- Two senses of “suit”: litigation vs. clothing
- Let’s represent these two senses using the embeddings $\vec{s}_1$, $\vec{s}_2$.

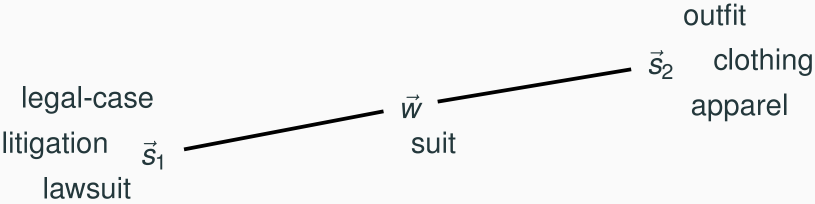
Why Some People Are Against single-embedding-per-word



- Two senses of “suit”: litigation vs. clothing
- Let’s represent these two senses using the embeddings $\vec{s}_1$, $\vec{s}_2$.
- Plausible approach: the embedding $\vec{w}$ of “suit” is

$$0.5(\vec{s}_1 + \vec{s}_2)$$

Why Some People Are Against single-embedding-per-word

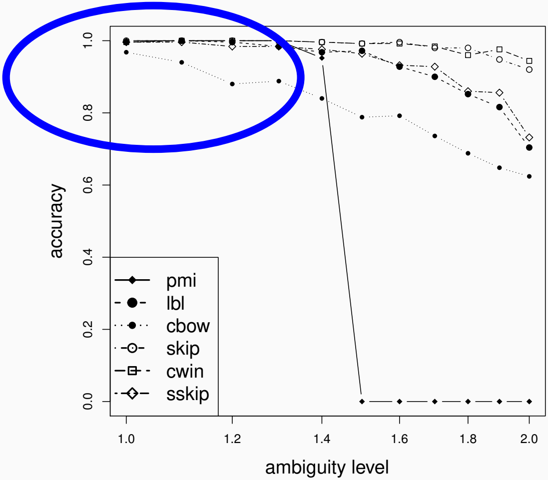


- Two senses of “suit”: litigation vs. clothing
- Let’s represent these two senses using the embeddings $\vec{s}_1$, $\vec{s}_2$.
- Plausible approach: the embedding $\vec{w}$ of “suit” is

$$0.5(\vec{s}_1 + \vec{s}_2)$$

- But $\vec{w}$ is not close to either senses (“litigation” / “clothing”)!

Ambiguity: Comparison of Six Embedding Models



Ambiguity: Summary

Ambiguity: Summary

- Our experiment shows:
  one-embedding-per-word is robust to some level of ambiguity

Ambiguity: Summary

- Our experiment shows:
  one-embedding-per-word is robust to some level of ambiguity
- Similar conclusion in (Li and Jurafsky (2015)): increasing word embedding dimensionality can do the job of multi sense embeddings

Ambiguity: Summary

- Our experiment shows:
  one-embedding-per-word is robust to some level of ambiguity
- Similar conclusion in (Li and Jurafsky (2015)): increasing word embedding dimensionality can do the job of multi sense embeddings
- However for skewed sense distributions, one-embedding-per-word is challenging

Challenge 2: Multifacetedness

- Words have large number of facets

Challenge 2: Multifacetedness

- Words have large number of facets
- Embedding should accurately represent all of them

Multifacetedness Grammar

1	$P(NF_n\|S)$	$=1/4$	
2	$P(AF_a\|S)$	$=1/4$	
3	$P(NM_n\|S)$	$=1/4$	
4	$P(AM_f\|S)$	$=1/4$	
5	$P(n_i\|N)$	$=1/5$	$0 \leq i \leq 4$
6	$P(a_i\|A)$	$=1/5$	$0 \leq i \leq 4$
7	$P(x_i^{\text{nf}} U_i^{\text{nf}}\|F_n)$	$=1/5$	$0 \leq i \leq 4$
8	$P(f\|U_i^{\text{nf}})$	$=1/2$	
9	$P(\mu(U_i^{\text{nf}})\|U_i^{\text{nf}})$	$=1/2$	
10	$P(x_i^{\text{af}} U_i^{\text{af}}\|F_a)$	$=1/5$	$0 \leq i \leq 4$
11	$P(f\|U_i^{\text{af}})$	$=1/2$	
12	$P(\mu(U_i^{\text{af}})\|U_i^{\text{af}})$	$=1/2$	
13	$P(x_i^{\text{nm}} U_i^{\text{nm}}\|M_n)$	$=1/5$	$0 \leq i \leq 4$
14	$P(m\|U_i^{\text{nm}})$	$=1/2$	
15	$P(\mu(U_i^{\text{nm}})\|U_i^{\text{nm}})$	$=1/2$	
16	$P(x_i^{\text{am}} U_i^{\text{am}}\|M_f)$	$=1/5$	$0 \leq i \leq 4$
17	$P(m\|U_i^{\text{am}})$	$=1/2$	
18	$P(\mu(U_i^{\text{am}})\|U_i^{\text{am}})$	$=1/2$	

This grammar generates nouns ($x_i^{\text{n}\cdot}$) and adjectives ($x_i^{\text{a}\cdot}$)

Multifacetedness Grammar

1	$P(NF_n \mid S)$	$= 1/4$	
2	$P(AF_a \mid S)$	$= 1/4$	
3	$P(NM_n \mid S)$	$= 1/4$	
4	$P(AM_f \mid S)$	$= 1/4$	
5	$P(n_i \mid N)$	$= 1/5$	$0 \leq i \leq 4$
6	$P(a_i \mid A)$	$= 1/5$	$0 \leq i \leq 4$
7	$P(x_i^{\text{nf}} U_i^{\text{nf}} \mid F_n)$	$= 1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\text{nf}})$	$= 1/2$	
9	$P(\mu(U_i^{\text{nf}}) \mid U_i^{\text{nf}})$	$= 1/2$	
10	$P(x_i^{\text{af}} U_i^{\text{af}} \mid F_a)$	$= 1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\text{af}})$	$= 1/2$	
12	$P(\mu(U_i^{\text{af}}) \mid U_i^{\text{af}})$	$= 1/2$	
13	$P(x_i^{\text{nm}} U_i^{\text{nm}} \mid M_n)$	$= 1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\text{nm}})$	$= 1/2$	
15	$P(\mu(U_i^{\text{nm}}) \mid U_i^{\text{nm}})$	$= 1/2$	
16	$P(x_i^{\text{am}} U_i^{\text{am}} \mid M_f)$	$= 1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\text{am}})$	$= 1/2$	
18	$P(\mu(U_i^{\text{am}}) \mid U_i^{\text{am}})$	$= 1/2$	

This grammar generates nouns ($x_i^{\text{n}\cdot}$) and adjectives ($x_i^{\text{a}\cdot}$)

Multifacetedness Grammar

1	$P(NF_n \mid S)$	$=1/4$	
2	$P(AF_a \mid S)$	$=1/4$	
3	$P(NM_n \mid S)$	$=1/4$	
4	$P(AM_f \mid S)$	$=1/4$	
5	$P(n_i \mid N)$	$=1/5$	$0 \leq i \leq 4$
6	$P(a_i \mid A)$	$=1/5$	$0 \leq i \leq 4$
7	$P(x_i^{\text{nf}} U_i^{\text{nf}} \mid F_n)$	$=1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\text{nf}})$	$=1/2$	
9	$P(\mu(U_i^{\text{nf}}) \mid U_i^{\text{nf}})$	$=1/2$	
10	$P(x_i^{\text{af}} U_i^{\text{af}} \mid F_a)$	$=1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\text{af}})$	$=1/2$	
12	$P(\mu(U_i^{\text{af}}) \mid U_i^{\text{af}})$	$=1/2$	
13	$P(x_i^{\text{nm}} U_i^{\text{nm}} \mid M_n)$	$=1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\text{nm}})$	$=1/2$	
15	$P(\mu(U_i^{\text{nm}}) \mid U_i^{\text{nm}})$	$=1/2$	
16	$P(x_i^{\text{am}} U_i^{\text{am}} \mid M_f)$	$=1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\text{am}})$	$=1/2$	
18	$P(\mu(U_i^{\text{am}}) \mid U_i^{\text{am}})$	$=1/2$	

with masculine (x_i^{m}) and feminine (x_i^{f}) gender

Multifacetedness Grammar

1	$P(NF_n \mid S)$	$= 1/4$	
2	$P(AF_a \mid S)$	$= 1/4$	
3	$P(NM_n \mid S)$	$= 1/4$	
4	$P(AM_f \mid S)$	$= 1/4$	
5	$P(n_i \mid N)$	$= 1/5$	$0 \leq i \leq 4$
6	$P(a_i \mid A)$	$= 1/5$	$0 \leq i \leq 4$
7	$P(x_i^{\text{nf}} U_i^{\text{nf}} \mid F_n)$	$= 1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\text{nf}})$	$= 1/2$	
9	$P(\mu(U_i^{\text{nf}}) \mid U_i^{\text{nf}})$	$= 1/2$	
10	$P(x_i^{\text{af}} U_i^{\text{af}} \mid F_a)$	$= 1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\text{af}})$	$= 1/2$	
12	$P(\mu(U_i^{\text{af}}) \mid U_i^{\text{af}})$	$= 1/2$	
13	$P(x_i^{\text{nm}} U_i^{\text{nm}} \mid M_n)$	$= 1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\text{nm}})$	$= 1/2$	
15	$P(\mu(U_i^{\text{nm}}) \mid U_i^{\text{nm}})$	$= 1/2$	
16	$P(x_i^{\text{am}} U_i^{\text{am}} \mid M_f)$	$= 1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\text{am}})$	$= 1/2$	
18	$P(\mu(U_i^{\text{am}}) \mid U_i^{\text{am}})$	$= 1/2$	

with masculine (x_i^{m}) and feminine (x_i^{f}) gender

Multifacetedness Grammar

1	$P(NF_n \mid S)$	$= 1/4$	
2	$P(AF_a \mid S)$	$= 1/4$	
3	$P(NM_n \mid S)$	$= 1/4$	
4	$P(AM_f \mid S)$	$= 1/4$	
5	$P(n_i \mid N)$	$= 1/5$	$0 \leq i \leq 4$
6	$P(a_i \mid A)$	$= 1/5$	$0 \leq i \leq 4$
7	$P(x_i^{\text{nf}} U_i^{\text{nf}} \mid F_n)$	$= 1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\text{nf}})$	$= 1/2$	
9	$P(\mu(U_i^{\text{nf}}) \mid U_i^{\text{nf}})$	$= 1/2$	
10	$P(x_i^{\text{af}} U_i^{\text{af}} \mid F_a)$	$= 1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\text{af}})$	$= 1/2$	
12	$P(\mu(U_i^{\text{af}}) \mid U_i^{\text{af}})$	$= 1/2$	
13	$P(x_i^{\text{nm}} U_i^{\text{nm}} \mid M_n)$	$= 1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\text{nm}})$	$= 1/2$	
15	$P(\mu(U_i^{\text{nm}}) \mid U_i^{\text{nm}})$	$= 1/2$	
16	$P(x_i^{\text{am}} U_i^{\text{am}} \mid M_f)$	$= 1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\text{am}})$	$= 1/2$	
18	$P(\mu(U_i^{\text{am}}) \mid U_i^{\text{am}})$	$= 1/2$	

Function μ maps each U to one of the morphological paradigms $\{u_0 \ldots u_4\}$

Multifacetedness: Example Sentences

The grammar generates three-word sentences

Multifacetedness: Example Sentences

The grammar generates three-word sentences

$a_1\ x_0^{\mathrm{af}}\ u_1$ $a_0\ x_0^{\mathrm{af}}\ f$	x_0^{af} : a feminine adjective with paradigm u_1
$a_4\ x_1^{\mathrm{af}}\ u_2$ $a_3\ x_1^{\mathrm{af}}\ f$	x_1^{af} : a feminine adjective with paradigm u_2
$n_3\ x_3^{\mathrm{nf}}\ u_3$ $n_1\ x_3^{\mathrm{nf}}\ f$	x_3^{nf} : a feminine noun with paradigm u_3
$n_3\ x_2^{\mathrm{nm}}\ u_1$ $n_2\ x_2^{\mathrm{nm}}\ m$	x_2^{nm} : a masculine noun with paradigm u_1

Multifacetedness: Experiment

Multifacetedness: Experiment

- SVM classification:
  “is this word feminine or masculine?”

Multifacetedness: Experiment

- SVM classification:
  “is this word feminine or masculine?”
- Training on the “nouns”, predict the gender of “adjectives”

Multifacetedness: Experiment

- SVM classification:
  "is this word feminine or masculine?"
- Training on the "nouns", predict the gender of "adjectives"
- 10 trials of experiments with different paradigm assignments

Multifacetedness: Results

- No single error in the classification
- Representation of gender facet by all six embedding models is perfect

Multifacetedness: Example Sentences

The grammar generates three-word sentences

$a_1\ x_0^{\mathrm{af}}\ u_1$ $a_0\ x_0^{\mathrm{af}}\ f$	x_0^{af} : a feminine adjective with paradigm u_1
$a_4\ x_1^{\mathrm{af}}\ u_2$ $a_3\ x_1^{\mathrm{af}}\ f$	x_1^{af} : a feminine adjective with paradigm u_2
$n_3\ x_3^{\mathrm{nf}}\ u_3$ $n_1\ x_3^{\mathrm{nf}}\ f$	x_3^{nf} : a feminine noun with paradigm u_3
$n_3\ x_2^{\mathrm{nm}}\ u_1$ $n_2\ x_2^{\mathrm{nm}}\ m$	x_2^{nm} : a masculine noun with paradigm u_1

Multifacetedness: Results

- No single error in the classification
- Representation of gender facet by all six embedding models is perfect
- But what if we used fullspace similarity?

Multifacetedness: Similarity/analogy Experiment

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
    - assign to each adjective, the gender of nearest neighbor in the train set

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
  - assign to each adjective, the gender of nearest neighbor in the train set
- Analogy evaluation:
  - randomly form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
  - assign to each adjective, the gender of nearest neighbor in the train set
- Analogy evaluation:
  - randomly form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$
  - if the nearest neighbor of $\vec{s}$ is feminine, search is successful

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
  - assign to each adjective, the gender of nearest neighbor in the train set
- Analogy evaluation:
  - randomly form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$
  - if the nearest neighbor of $\vec{s}$ is feminine, search is successful
- Error rate results:

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
  - assign to each adjective, the gender of nearest neighbor in the train set
- Analogy evaluation:
  - randomly form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$
  - if the nearest neighbor of $\vec{s}$ is feminine, search is successful
- Error rate results:
  - similarity: $\sim 20\%$
  - analogy: $\sim 15\%$

Multifacetedness: Similarity/analogy Experiment

- Similarity evaluation:
  - assign to each adjective, the gender of nearest neighbor in the train set
- Analogy evaluation:
  - randomly form analogies like: $\vec{s} = \vec{x}_i^{\mathrm{nf}} - \vec{x}_j^{\mathrm{nm}} + \vec{x}_k^{\mathrm{am}}$
  - if the nearest neighbor of $\vec{s}$ is feminine, search is successful
- Error rate results:
  - similarity: $\sim$ 20%
  - analogy: $\sim$ 15%
  - classification: 0%

Evaluation of Multifacetedness: Summary

Evaluation of Multifacetedness: Summary

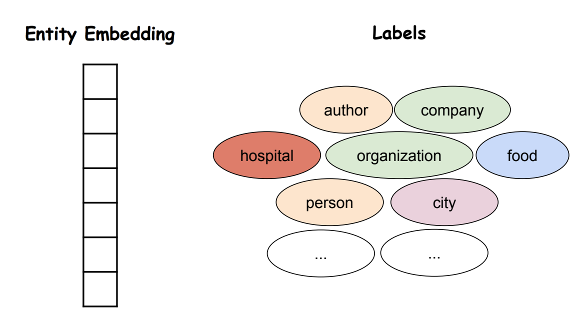
- Fullspace similarity is misleading

Evaluation of Multifacetedness: Summary

- Fullspace similarity is misleading
- Subspace similarity works well

Extrinsic Evaluation: Entity Typing

Task


Entity Embedding
Labels
author
company
hospital
organization
food
person
city
...
...

Entity Typing: Experiment

Entity Typing: Experiment

- Two types of classifiers:
  - fullspace similarity based: KNN
  - subspace similarity based: MLP

Results

	MLP	KNN
PPMI	61.6	44.0
LBL	63.5	51.7
CBOW	63.0	53.5
CWIN	66.1	53.0
SKIP	64.5	57.1
SSKIP	66.2	52.8

- MLP $>>$ KNN (i.e., subspace sim $>>$ fullspace sim)

Results

	MLP	KNN
PPMI	61.6	44.0
LBL	63.5	51.7
CBOW	63.0	53.5
CWIN	66.1	53.0
SKIP	64.5	57.1
SSKIP	66.2	52.8

- MLP $>>$ KNN (i.e., subspace sim $>>$ fullspace sim)
- Fullspace sim is misleading:

Results

	MLP	KNN
PPMI	61.6	44.0
LBL	63.5	51.7
CBOW	63.0	53.5
CWIN	66.1	53.0
SKIP	64.5	57.1
SSKIP	66.2	52.8

- MLP $>>$ KNN
  (i.e., subspace sim $>>$ fullspace sim)
- Fullspace sim is misleading:
  - fullspace sim incorrectly suggests:
    SSKIP is worse than SKIP

Results

	MLP	KNN
PPMI	61.6	44.0
LBL	63.5	51.7
CBOW	63.0	53.5
CWIN	66.1	53.0
SKIP	64.5	57.1
SSKIP	66.2	52.8

- MLP $>>$ KNN
  (i.e., subspace sim $>>$ fullspace sim)
- Fullspace sim is misleading:
  - fullspace sim incorrectly suggests:
    SSKIP is worse than SKIP
  - subspace sim shows that in reality:
    SSKIP is better than SKIP

Entity Typing: Results (2)

	all entities		frequent entities		rare entities	
	MLP	KNN	MLP	KNN	MLP	KNN
PPMI	61.6	44.0	69.2	63.8	43.0	28.5
LBL	63.5	51.7	72.7	66.4	44.1	32.8
CBOW	63.0	53.5	71.7	69.4	39.1	29.9
CWIN	66.1	53.0	73.5	68.6	46.8	31.4
SKIP	64.5	57.1	69.9	71.5	49.8	34.0
SSKIP	66.2	52.8	73.9	68.5	45.5	31.4

Entity Typing: Results (2)

	all entities		frequent entities		rare entities	
	MLP	KNN	MLP	KNN	MLP	KNN
PPMI	61.6	44.0	69.2	63.8	43.0	28.5
LBL	63.5	51.7	72.7	66.4	44.1	32.8
CBOW	63.0	53.5	71.7	69.4	39.1	29.9
CWIN	66.1	53.0	73.5	68.6	46.8	31.4
SKIP	64.5	57.1	69.9	71.5	49.8	34.0
SSKIP	66.2	52.8	73.9	68.5	45.5	31.4

- Spearman's rho between subspace sim (MLP) and fullspace sim (KNN) is: 0.03 for frequent and 0.75 for rare

Entity Typing: Results (2)

	all entities		frequent entities		rare entities	
	MLP	KNN	MLP	KNN	MLP	KNN
PPMI	61.6	44.0	69.2	63.8	43.0	28.5
LBL	63.5	51.7	72.7	66.4	44.1	32.8
CBOW	63.0	53.5	71.7	69.4	39.1	29.9
CWIN	66.1	53.0	73.5	68.6	46.8	31.4
SKIP	64.5	57.1	69.9	71.5	49.8	34.0
SSKIP	66.2	52.8	73.9	68.5	45.5	31.4

- Spearman's rho between subspace sim (MLP) and fullspace sim (KNN) is: 0.03 for frequent and 0.75 for rare
- What does the correlation difference in frequent and rare entities tell us?

Summary

- Intrinsic evaluation is needed
- Fullspace similarity is limited and misleading
- Subspace similarity is necessary for a good evaluation
- Artificial corpora can help to evaluate and understand

Thank you!

yadollah@cis.lmu.de