Intrinsic Subspace Evaluation of Word Embedding Representations

December 2015

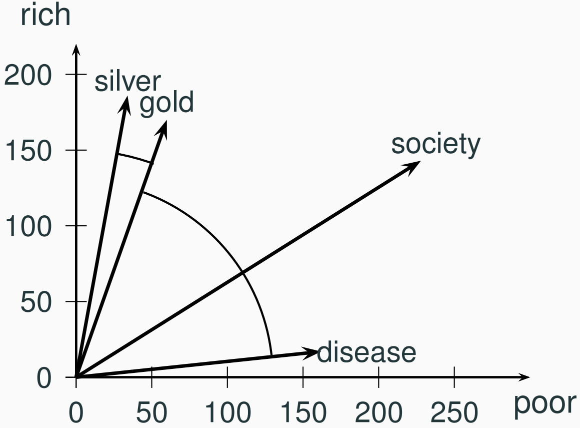
LMU Munich
CIS

Outline

Word Embeddings Evaluation

Word Embeddings Evaluation

Representation of Words in Vector space model



Word Embeddings

- Word embedding: real-valued vector representing all the facets of the word

- Are these methods proper for the evaluation?

Word Embeddings

- Word embedding: real-valued vector representing all the facets of the word
- Evaluation: how much of these facets are represented accurately and consistently.

- Are these methods proper for the evaluation?

Word Embeddings

- Word embedding: real-valued vector representing all the facets of the word
- Evaluation: how much of these facets are represented accurately and consistently.
- Current methods of evaluation:
    - Extrinsic tasks
    - Intrinsic similarity-based datasets
- Are these methods proper for the evaluation?

Our Evaluation Method

- Objective: how much of word facets are represented accurately and consistently.

Our Evaluation Method

- Objective: how much of word facets are represented accurately and consistently.
- Our proposed method:
  - Identifying generic and fundamental criteria for embedding models that are important to represent facets of words accurately and consistently.

Our Evaluation Method

- Objective: how much of word facets are represented accurately and consistently.
- Our proposed method:
  - Identifying generic and fundamental criteria for embedding models that are important to represent facets of words accurately and consistently.
  - Developing corpus-based tests to asses the criteria using PCFG generated datasets

Our Evaluation Method

- Objective: how much of word facets are represented accurately and consistently.
- Our proposed method:
  - Identifying generic and fundamental criteria for embedding models that are important to represent facets of words accurately and consistently.
  - Developing corpus-based tests to asses the criteria using PCFG generated datasets
  - Evaluating embedding models by supervised classification that needs to look at subspaces

Criteria for word embeddings

- Nonconflation

Criteria for word embeddings

- Nonconflation
- Robustness against sparseness

Criteria for word embeddings

- Nonconflation
- Robustness against sparseness
- Robustness against ambiguity

Criteria for word embeddings

- Nonconflation
- Robustness against sparseness
- Robustness against ambiguity
- Accurate and consistent representation of multifacetedness

Criteria for word embeddings

- Robustness against ambiguity
- Accurate and consistent representation of multifacetedness

Robustness against ambiguity

- Most words are ambiguous.

Robustness against ambiguity

- Important question:
    - for a word with k senses, Should we have k vectors? Or just 1 vector?

*1*vector-$k$senses: 1 embedding to represent two senses?

1 vector-k senses: 1 embedding to represent two senses?

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

*1*vector-*k*senses: 1 embedding to represent two senses?

outfit

clothing

legal-case

apparel

litigation

lawsuit

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

*1*vector-*k*senses: 1 embedding to represent two senses?

outfit

clothing

legal-case

apparel

litigation

lawsuit

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

*1*vector-$k$senses: 1 embedding to represent two senses?

outfit

$\vec{s}_2$ clothing

apparel

legal-case

litigation $\vec{s}_1$

lawsuit

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

*1*vector-*k*senses: 1 embedding to represent two senses?

outfit

$\vec{s}_2$ clothing

apparel

legal-case

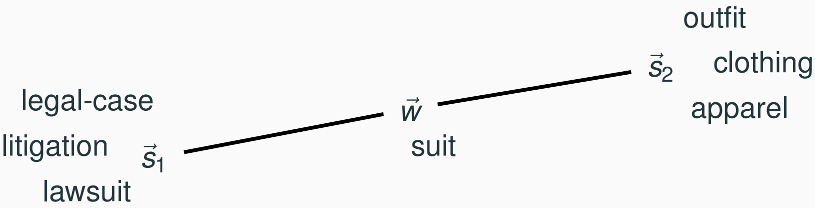
litigation $\vec{s}_1$

lawsuit

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

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

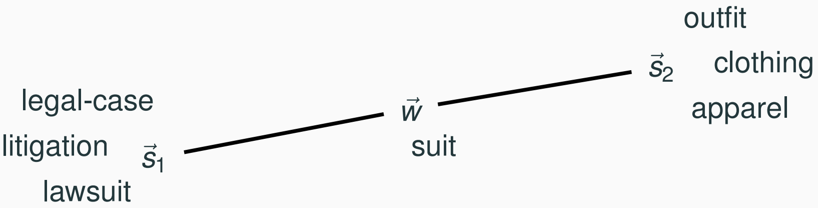
*1*vector-*k*senses: 1 embedding to represent two senses?



- 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

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

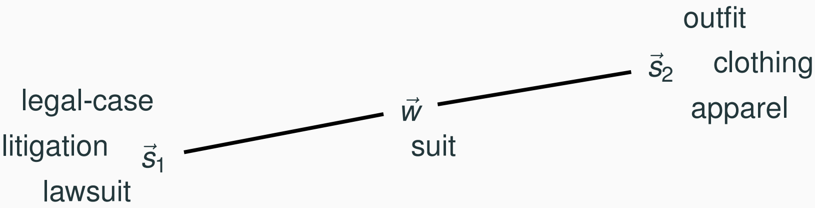
*1*vector-*k*senses: 1 embedding to represent two senses?



- 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

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

*1*vector-*k*senses: 1 embedding to represent two senses?

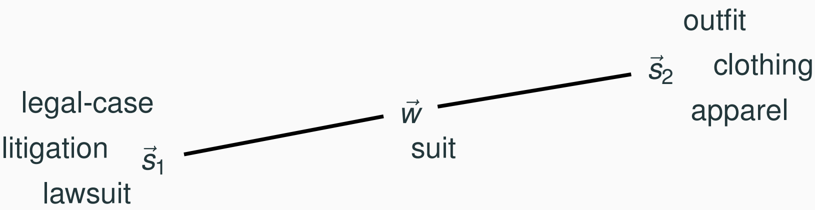


- Two senses of “suit”: litigation vs. clothing
- Let’s represent these two senses using the embeddings $\vec{s}_1$, $\vec{s}_2$.
- Plausible approach: the embedding $\vec{w}$ of “suit” is

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

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

*1*vector-*k*senses: 1 embedding to represent two senses?



- 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 sense ("litigation" / "clothing")!
- Does that mean that *1*vector-*k*senses is a bad idea?

Embeddings: For which objectives?

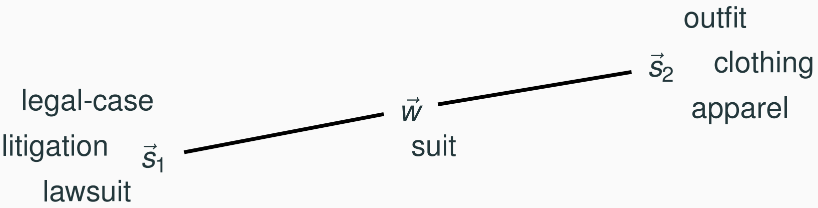
Embeddings: For which objectives?

- It is often assumed:
  - closeness in embedding space should model closeness in “linguistic” space.

Embeddings: For which objectives?

- It is often assumed:
  - closeness in embedding space should model closeness in “linguistic” space.
- Similarity evaluation datasets
  (MEN, WordSim, Stanford Rare Words etc)
  evaluate performance on this objective.

*1*vector-*k*senses: 1 embedding to represent two senses?



- 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 sense ("litigation" / "clothing")!
- Does that mean that *1*vector-*k*senses is a bad idea?

Embeddings: Remember our objective

Embeddings: Remember our objective

- Word embeddings should represent the facets of words accurately and consistently

Embeddings: Remember our objective

- Word embeddings should represent the facets of words accurately and consistently
- If we adopt this alternative objective,
  then maybe there is no problem with *1*vector-*k*senses?

Setup for ambiguity experiment

Setup for ambiguity experiment

- Define PCFG grammar

Setup for ambiguity experiment

- Define PCFG grammar
- PCFG models ambiguity in natural language

Setup for ambiguity experiment

- Define PCFG grammar
- PCFG models ambiguity in natural language
- Generate a corpus using the PCFG

Setup for ambiguity experiment

- Define PCFG grammar
- PCFG models ambiguity in natural language
- Generate a corpus using the PCFG
- Train embedding models on corpus:
  each will give us its trained word embeddings

Ambiguity grammar

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

Ambiguity grammar

$P(AV_1B\|S)$	=	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	$0 \leq i \leq 4$

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

Ambiguity grammar

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

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

Ambiguity grammar

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

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

Ambiguity grammar

$P(AV_1B\|S)$	=	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: $v_0 \ldots v_{49}$ only occur in A-B contexts.

Ambiguity grammar

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

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

Ambiguity grammar

$P(AV_1B\|S)$	=	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	$0 \leq i \leq 4$

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

Ambiguity grammar

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

Skewedness of ambiguity controlled by β.

Ambiguity: Experiment

Ambiguity: Experiment

- Learn embeddings from corpus

Ambiguity: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"

Ambiguity: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"
- One SVM for embeddings of each model

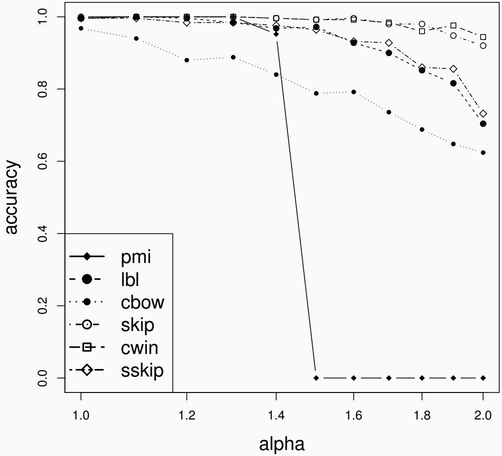
Ambiguity: Experiment

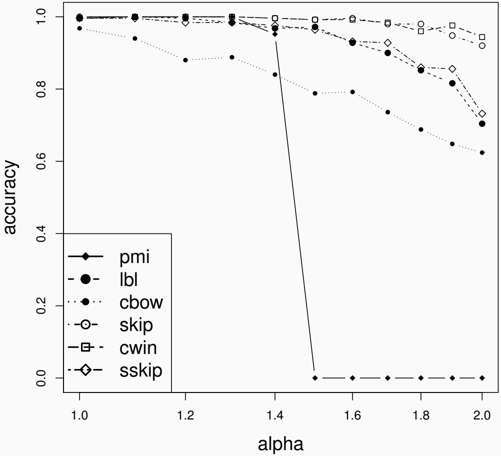
- Learn embeddings from corpus
- SVMs were trained for the binary classification task
  "can this word occur in an C-D context?"
- One SVM for embeddings of each model
- The test set consists of the five ambiguous words $w_0 \ldots w_4$.

Ambiguity: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"
- One SVM for embeddings of each model
- The test set consists of the five ambiguous words $w_0 \ldots w_4$.
- 50 trials of this experiment were run for each of eleven values of β: $\beta = 2^{-\alpha}$ where $\alpha \in \{1.0, 1.1, 1.2, \ldots, 2.0\}$.

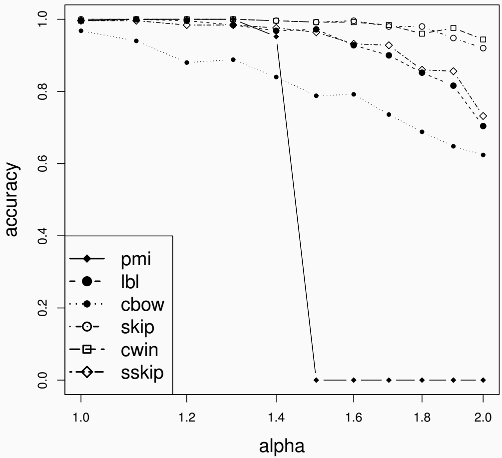
Ambiguity: Results





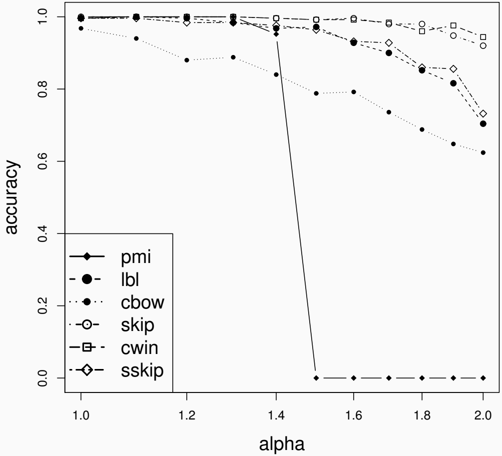
- x-axis: skewedness parameter β (1.0 $\approx$ 50/50, 2.0 $\approx$ 25/75)

Ambiguity: Results



- x-axis: skewedness parameter β (1.0 $\approx$ 50/50, 2.0 $\approx$ 25/75)
- Most embedding models degrade gracefully as skewedness increases.

Ambiguity: Results



- x-axis: skewedness parameter β (1.0 $\approx$ 50/50, 2.0 $\approx$ 25/75)
- Most embedding models degrade gracefully as skewedness increases.
- PPMI accuracy is 0 for *beta* $\geq$ 1.5.

Ambiguity: Summary

Ambiguity: Summary

- Premise of much work on ambiguity:
  *k*vectors-*k*senses: We need one embedding per sense.

Ambiguity: Summary

- Premise of much work on ambiguity:
  *k*vectors-*k*senses: We need one embedding per sense.
- But maybe we only need a single embedding per word
  i.e., *1*vector-*k*senses

Ambiguity: Summary

- Premise of much work on ambiguity:
  *k*vectors-*k*senses: We need one embedding per sense.
- But maybe we only need a single embedding per word
  i.e., *1*vector-*k*senses
- A single embedding will not fare well on analogy/similarity.

Ambiguity: Summary

- Premise of much work on ambiguity:
  *k*vectors-*k*senses: We need one embedding per sense.
- But maybe we only need a single embedding per word
  i.e., *1*vector-*k*senses
- A single embedding will not fare well on analogy/similarity.
- But it is not the word embedding objective.

Accurate and consistent representation of multifacetedness

- Words have a large number of facets

Accurate and consistent representation of multifacetedness

- Words have a large number of facets
- Each facet constitutes a small part of the overall information that a representation should capture about a word.

Setup for multifacetedness experiment

Setup for multifacetedness experiment

- Define PCFG grammar

Setup for multifacetedness experiment

- Define PCFG grammar
- PCFG models multifacetedness in natural language

Setup for multifacetedness experiment

- Define PCFG grammar
- PCFG models multifacetedness in natural language
- Generate a corpus using the PCFG

Setup for multifacetedness experiment

- Define PCFG grammar
- PCFG models multifacetedness in natural language
- Generate a corpus using the PCFG
- Train embedding models on the corpus:
  each gives us a set of word embeddings.

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	$P(x_i^{\mathrm{nf}} U_i^{\mathrm{nf}} \mid F_n)$	$= 1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\mathrm{nf}})$	$= 1/2$	
9	$P(\mu(U_i^{\mathrm{nf}}) \mid U_i^{\mathrm{nf}})$	$= 1/2$	
10	$P(x_i^{\mathrm{af}} U_i^{\mathrm{af}} \mid F_a)$	$= 1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\mathrm{af}})$	$= 1/2$	
12	$P(\mu(U_i^{\mathrm{af}}) \mid U_i^{\mathrm{af}})$	$= 1/2$	
13	$P(x_i^{\mathrm{nm}} U_i^{\mathrm{nm}} \mid M_n)$	$= 1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\mathrm{nm}})$	$= 1/2$	
15	$P(\mu(U_i^{\mathrm{nm}}) \mid U_i^{\mathrm{nm}})$	$= 1/2$	
16	$P(x_i^{\mathrm{am}} U_i^{\mathrm{am}} \mid M_f)$	$= 1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\mathrm{am}})$	$= 1/2$	
18	$P(\mu(U_i^{\mathrm{am}}) \mid U_i^{\mathrm{am}})$	$= 1/2$	

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

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	$P(x_i^{\mathrm{nf}} U_i^{\mathrm{nf}} \mid F_n)$	$= 1/5$	$0 \leq i \leq 4$
8	$P(f \mid U_i^{\mathrm{nf}})$	$= 1/2$	
9	$P(\mu(U_i^{\mathrm{nf}}) \mid U_i^{\mathrm{nf}})$	$= 1/2$	
10	$P(x_i^{\mathrm{af}} U_i^{\mathrm{af}} \mid F_a)$	$= 1/5$	$0 \leq i \leq 4$
11	$P(f \mid U_i^{\mathrm{af}})$	$= 1/2$	
12	$P(\mu(U_i^{\mathrm{af}}) \mid U_i^{\mathrm{af}})$	$= 1/2$	
13	$P(x_i^{\mathrm{nm}} U_i^{\mathrm{nm}} \mid M_n)$	$= 1/5$	$0 \leq i \leq 4$
14	$P(m \mid U_i^{\mathrm{nm}})$	$= 1/2$	
15	$P(\mu(U_i^{\mathrm{nm}}) \mid U_i^{\mathrm{nm}})$	$= 1/2$	
16	$P(x_i^{\mathrm{am}} U_i^{\mathrm{am}} \mid M_f)$	$= 1/5$	$0 \leq i \leq 4$
17	$P(m \mid U_i^{\mathrm{am}})$	$= 1/2$	
18	$P(\mu(U_i^{\mathrm{am}}) \mid U_i^{\mathrm{am}})$	$= 1/2$	

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

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	$P(\mu(U_i^{\text{am}})\|U_i^{\text{am}})$	$=1/2$	

with masculine ($x_i{}^{\text{m}}$) and feminine ($x_i{}^{\text{f}}$) 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 paradigms $\{u_0 \ldots u_4\}$

Multifacetedness: Experiment

Multifacetedness: Experiment

- Learn embeddings from corpus

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task
  "can this word occur in an C-D context?"
- One SVM for embeddings of each model

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task
  “can this word occur in an C-D context?”
- One SVM for embeddings of each model
- Training on feminine and masculine “nouns”, the task is to predict gender of “adjectives”.

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task “can this word occur in an C-D context?”
- One SVM for embeddings of each model
- Training on feminine and masculine “nouns”, the task is to predict gender of “adjectives”.
- 10 trials of experiments with different paradigm assignments.

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"
- One SVM for embeddings of each model
- Training on feminine and masculine "nouns", the task is to predict gender of "adjectives".
- 10 trials of experiments with different paradigm assignments.
- Results: No single error in the classification: All models could predict the gender 100% accurate.

Multifacetedness: Experiment

- Learn embeddings from corpus
- SVMs were trained for the binary classification task "can this word occur in an C-D context?"
- One SVM for embeddings of each model
- Training on feminine and masculine "nouns", the task is to predict gender of "adjectives".
- 10 trials of experiments with different paradigm assignments.
- Results: No single error in the classification: All models could predict the gender 100% accurate.
- But what if we used similarity/analogy to check this criterion?

Multifacetedness: similarity/analogy experiment

Multifacetedness: similarity/analogy experiment

- But what if we used similarity/analogy to check this criterion?
- similarity evaluation: assign to each adjective, the gender of nearest neighbour in the train set.
- analogy evaluation: randomly select triples to form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$ the search is successful if the nearest neighbor is feminine.

Multifacetedness: similarity/analogy experiment

- But what if we used similarity/analogy to check this criterion?
- similarity evaluation: assign to each adjective, the gender of nearest neighbour in the train set.
- analogy evaluation: randomly select triples to form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$ the search is successful if the nearest neighbor is feminine.
- We did this evaluation on the same test set as before.

Multifacetedness: similarity/analogy experiment

- But what if we used similarity/analogy to check this criterion?
- similarity evaluation: assign to each adjective, the gender of nearest neighbour in the train set.
- analogy evaluation: randomly select triples to form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$ the search is successful if the nearest neighbor is feminine.
- We did this evaluation on the same test set as before.
- Results: error rates of around 20% similarity and around 15% for analogy.

Multifacetedness: similarity/analogy experiment

- But what if we used similarity/analogy to check this criterion?
- similarity evaluation: assign to each adjective, the gender of nearest neighbour in the train set.
- analogy evaluation: randomly select triples to form analogies like: $\vec{s} = \vec{x}_i^{\text{nf}} - \vec{x}_j^{\text{nm}} + \vec{x}_k^{\text{am}}$ the search is successful if the nearest neighbor is feminine.
- We did this evaluation on the same test set as before.
- Results: error rates of around 20% similarity and around 15% for analogy.
- Note that we got 0% error for SVM classification to find genders.

Multifacetedness: similarity/analogy experiment

- But what if we used similarity/analogy to check this criterion?
- similarity evaluation: assign to each adjective, the gender of nearest neighbour in the train set.
- analogy evaluation: randomly select triples to form analogies like: $\vec{s} = \vec{x}_i^{\mathrm{nf}} - \vec{x}_j^{\mathrm{nm}} + \vec{x}_k^{\mathrm{am}}$ the search is successful if the nearest neighbor is feminine.
- We did this evaluation on the same test set as before.
- Results: error rates of around 20% similarity and around 15% for analogy.
- Note that we got 0% error for SVM classification to find genders.
- Full-space similarity cannot determine the presence of a low entropy facet accurately and consistently.