Discretized Word Representations Meet Knowledge Graphs

Gábor Berend

09/11/2017
GraphNLP meetup

Semantics

- What is in the meaning of a word?

Computational Semantics

- What is in the meaning of a word?
- How shall we represent it for computers?

Computational Semantics

- What is in the meaning of a word?
- How shall we represent it for computers?
- Distributional hypothesis
  - Words with similar meaning tend to be present in similar contexts
    - The dog chased the cat.
    - The predator pursued the prey.
    - A cheetah catched an antilop.

Computational Semantics

- What is in the meaning of a word?
- How shall we represent it for computers?
- Distributional hypothesis
  - Words with similar meaning tend to be present in similar contexts
    - The dog chased the cat.
    - The predator pursued the prey.
    - A cheetah catched an antilop.
  - > ½ a century old theory

Computational Semantics

- What is in the meaning of a word?
- How shall we represent it for computers?
- Distributional hypothesis
  - Words with similar meaning tend to be present in similar contexts
    - The dog chased the cat.
    - The predator pursued the prey.
    - A cheetah catched an antilop.
  - > ½ a century old theory
    - < ½ decade extreme enthusiasm around it

Competing paradigms

- Increased awareness since 2013 (word2vec)

Competing paradigms

- Increased awareness since 2013 (word2vec)
  - 2014

***Don't count, predict!* A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**

Competing paradigms

- Increased awareness since 2013 (word2vec)
  - 2014

Don't count, predict! **A systematic comparison of context-counting vs. context-predicting semantic vectors**

Marco Baroni and **Georgiana Dinu** and **Germán Kruszewski**

  - 2015

Rehabilitation of Count-based Models for Word Vector Representations

Rémi Lebret[1,2] and Ronan Collobert[1]

The (implicit) goal of word embeddings

- Map word forms to such vectors that they reflect their co-occurrence statistics

The (implicit) goal of word embeddings

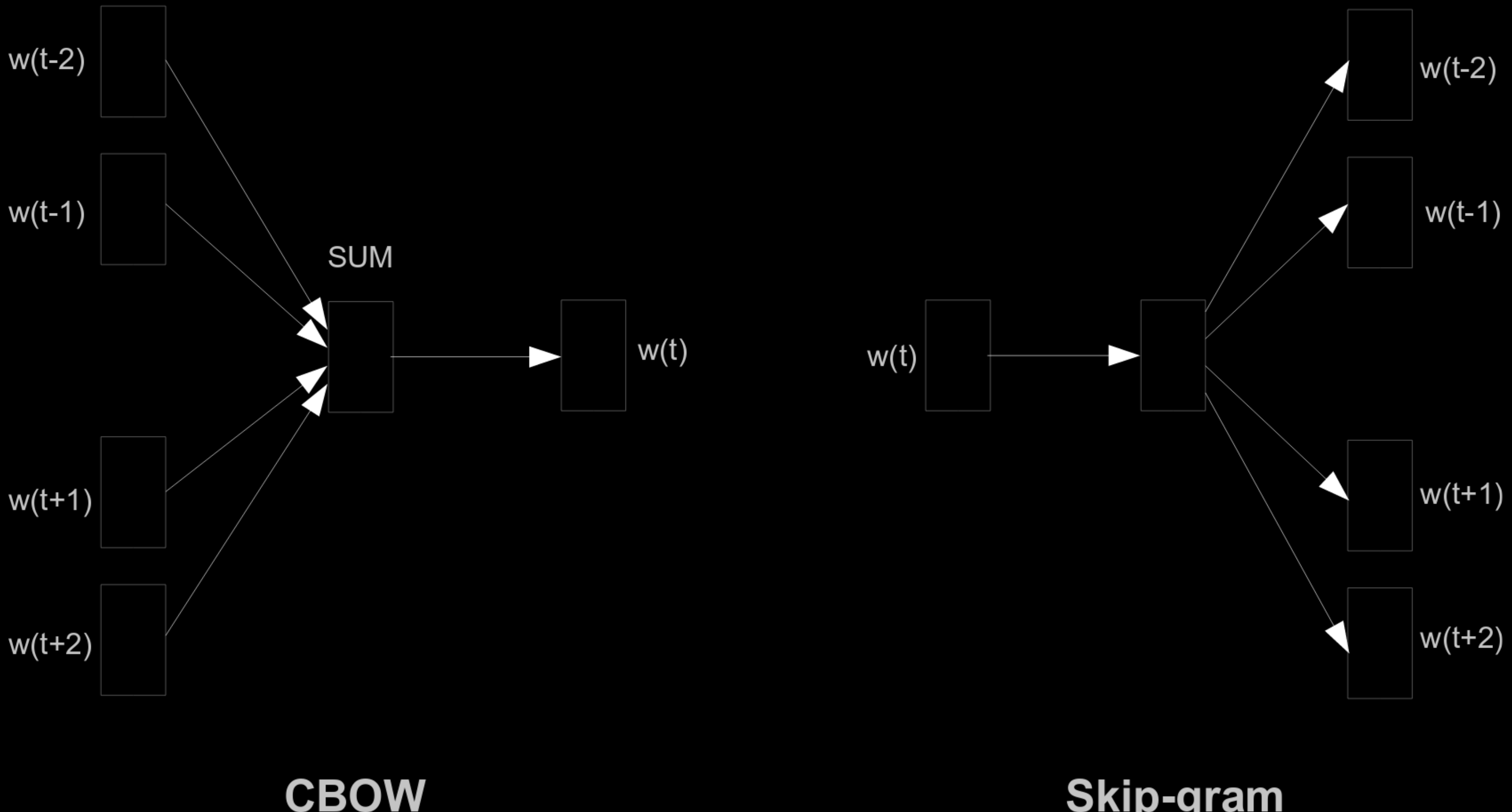
- Map word forms to such vectors that they reflect their co-occurrence statistics
  - Let vectors of words with (dis)similar meaning point to (dis)similar directions

The (implicit) goal of word embeddings

- Map word forms to such vectors that they reflect their co-occurrence statistics
  - Let vectors of words with (dis)similar meaning point to (dis)similar directions

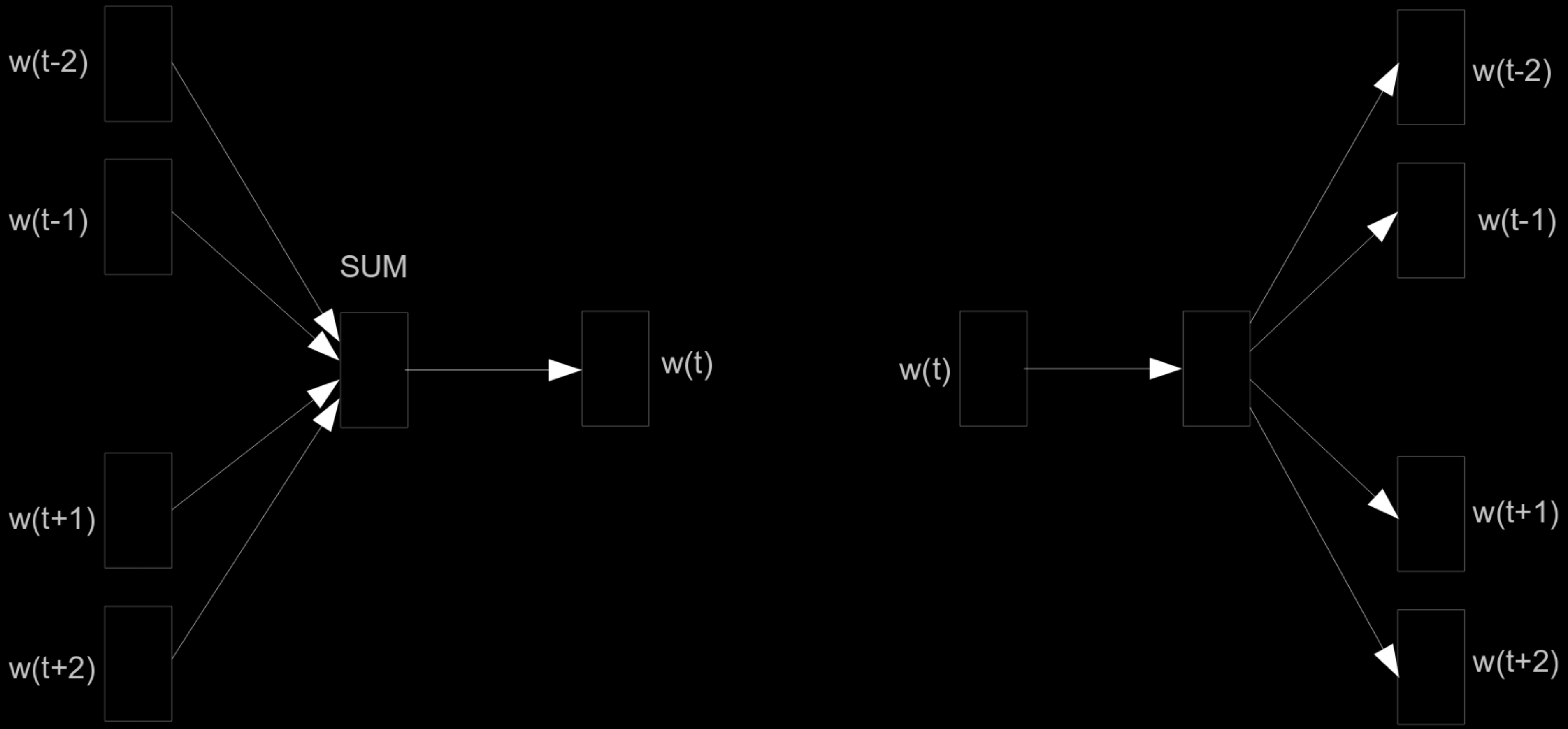
$$\overrightarrow{dog}^T \overrightarrow{cat} \gg \overrightarrow{dog}^T \overrightarrow{train}$$

word2vec variants



word2vec variants

quick brown X jumps over

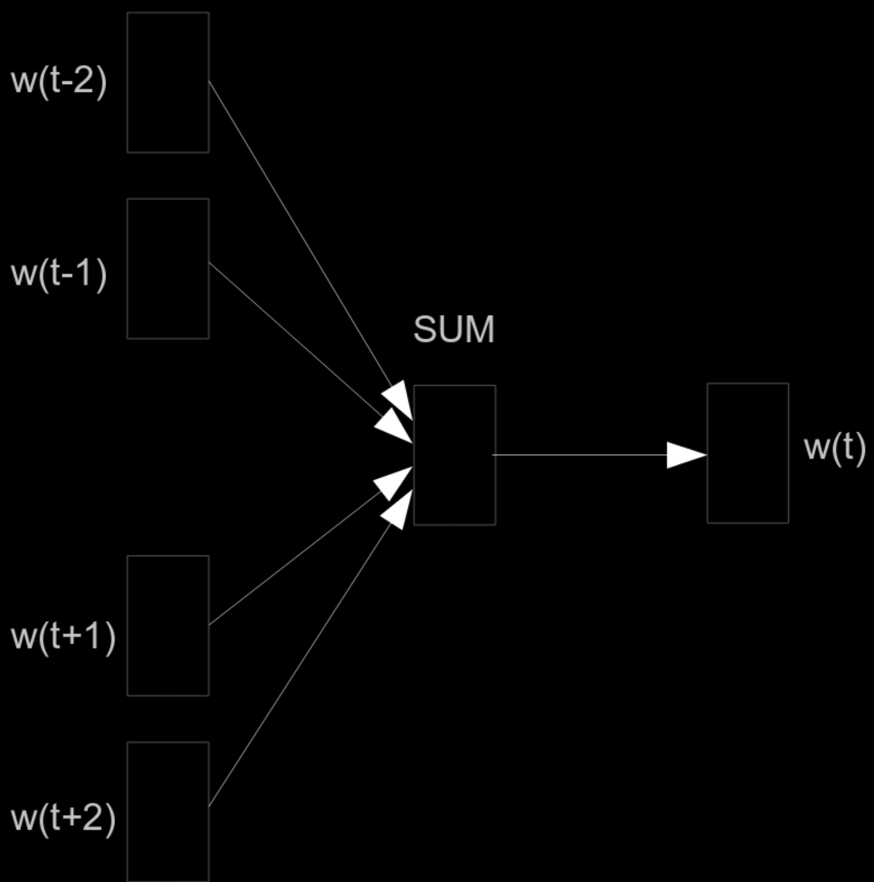


CBOW **Skip-gram**

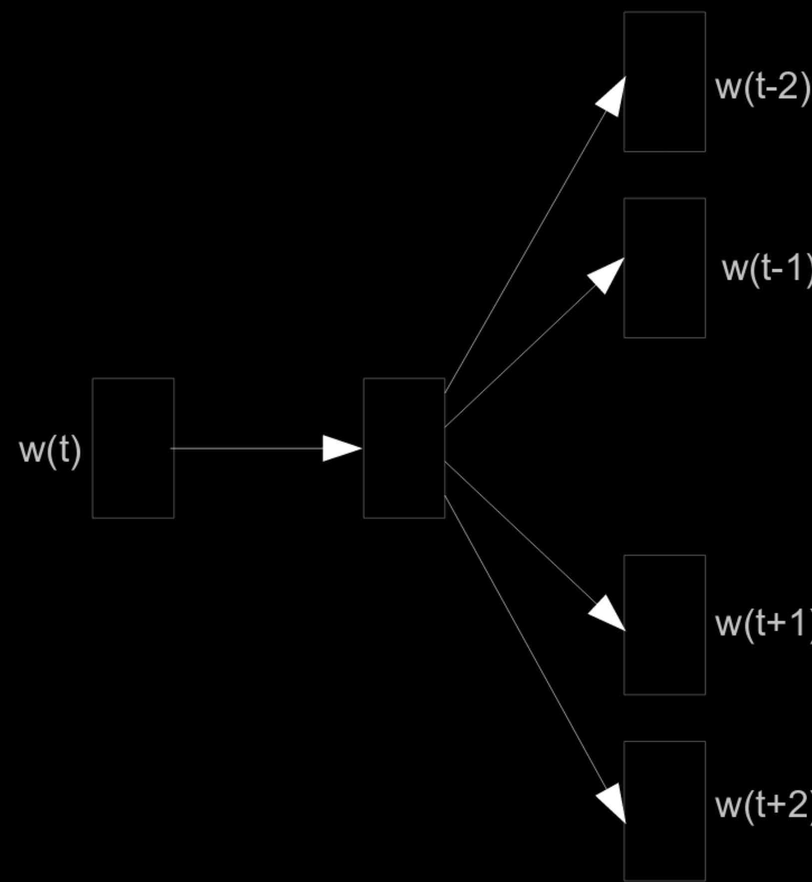
word2vec variants

quick brown X jumps over

U V fox Y Z



CBOW



Skip-gram

The goal of word2vec

- The NN view: given an input word **x** ‘predict’ an output word which fits in its context

$$y(x) = softmax(V(W 1_x))$$

  - The more similar two input vectors, the more similar their predictions tend to be

Continuous word representations

apple [1 0 0 0 … 0 0 0 0 0 … 0] → [3.2 -1.5]

...

banana [0 0 0 0 … 1 0 0 0 0 … 0] → [2.8 -1.6]

...

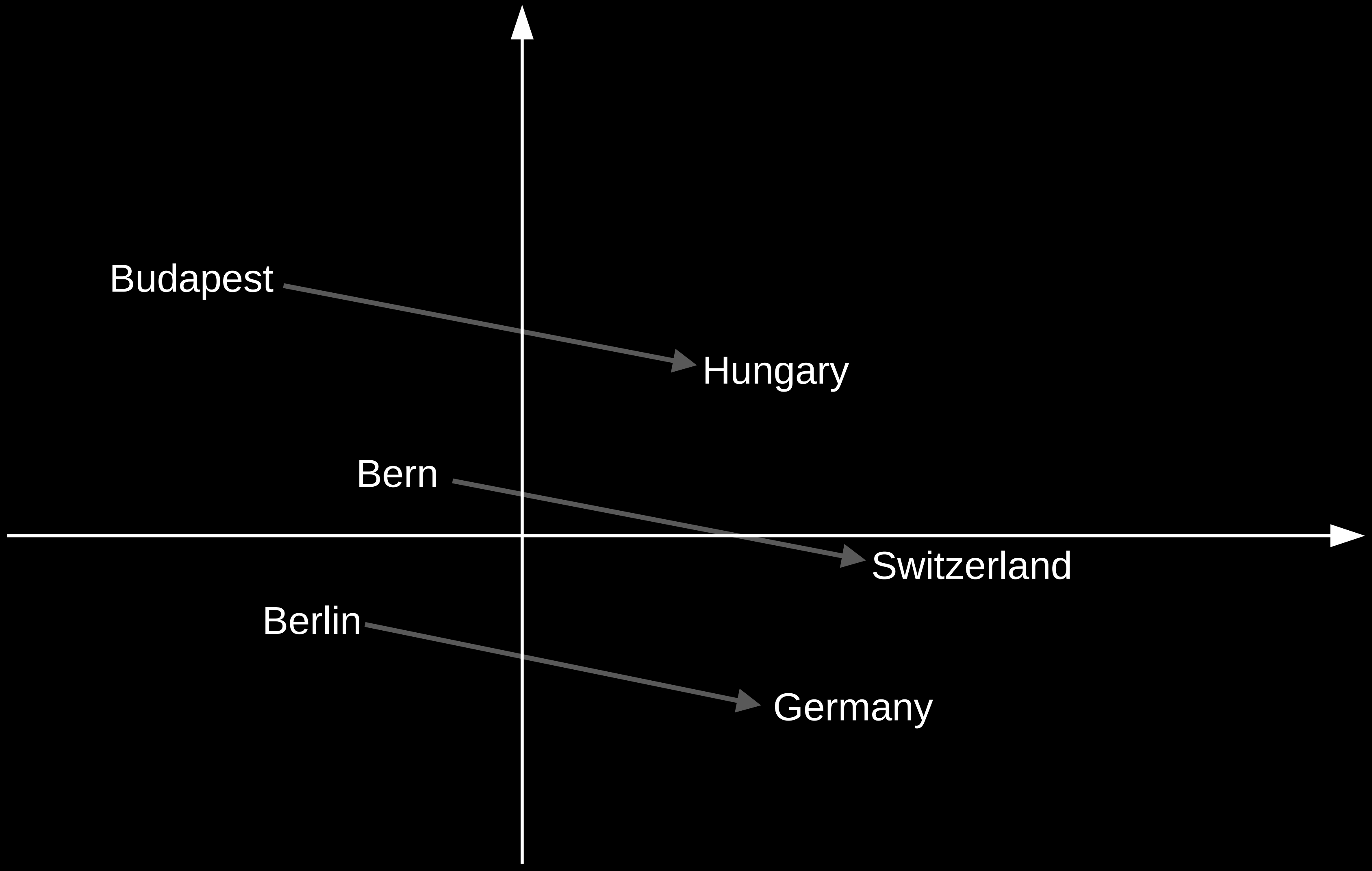
door [0 0 0 0 … 0 0 1 0 0 … 0] → [-1.1 12.6]

…

zebra [0 0 0 0 … 0 0 0 0 0 … 1] → [0.8 0.5]

Word analogies

- a:b::c:?



RepEval 2016

Analysis Track

- **Problems With Evaluation of Word Embeddings Using Word Similarity Tasks** [pdf]
  Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer
- **Intrinsic Evaluations of Word Embeddings: What Can We Do Better?** [pdf]
  Anna Gladkova, Aleksandr Drozd
- **Issues in Evaluating Semantic Spaces Using Word Analogies** [pdf]
  Tal Linzen
- **Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance** [pdf]
  Billy Chiu, Anna Korhonen, Sampo Pyysalo
- **A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models** [pdf]
  Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, David Weir

RepEval 2016

Analysis Track

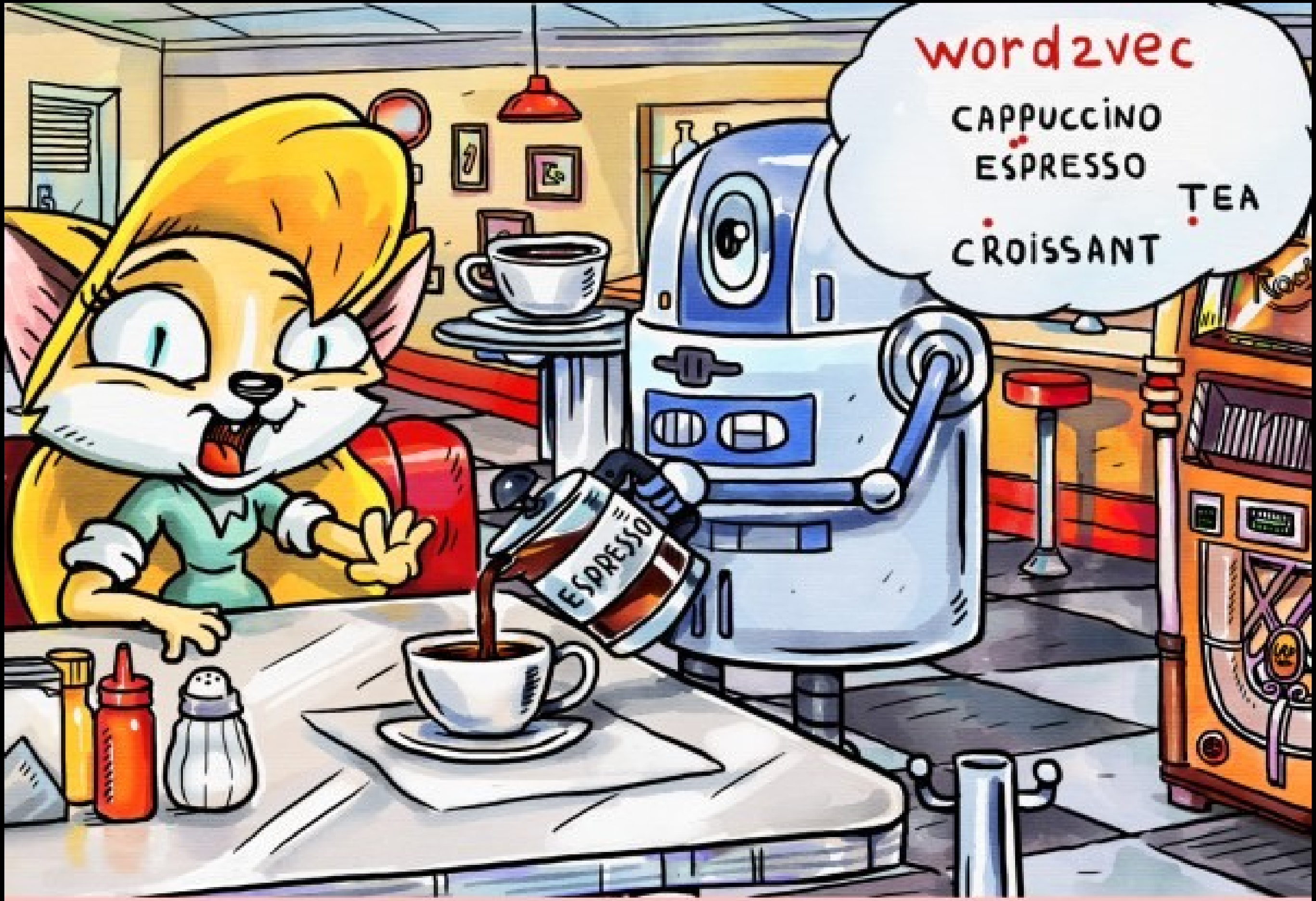
- **Problems With Evaluation of Word Embeddings Using Word Similarity Tasks** [pdf]
  Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, Chris Dyer

- **Intrinsic Evaluations of Word Embeddings: What Can We Do Better?** [pdf]
  Anna Gladkova, Aleksandr Drozd

- **Issues in Evaluating Semantic Spaces Using Word Analogies** [pdf]
  Tal Linzen

- **Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance** [pdf]
  Billy Chiu, Anna Korhonen, Sampo Pyysalo

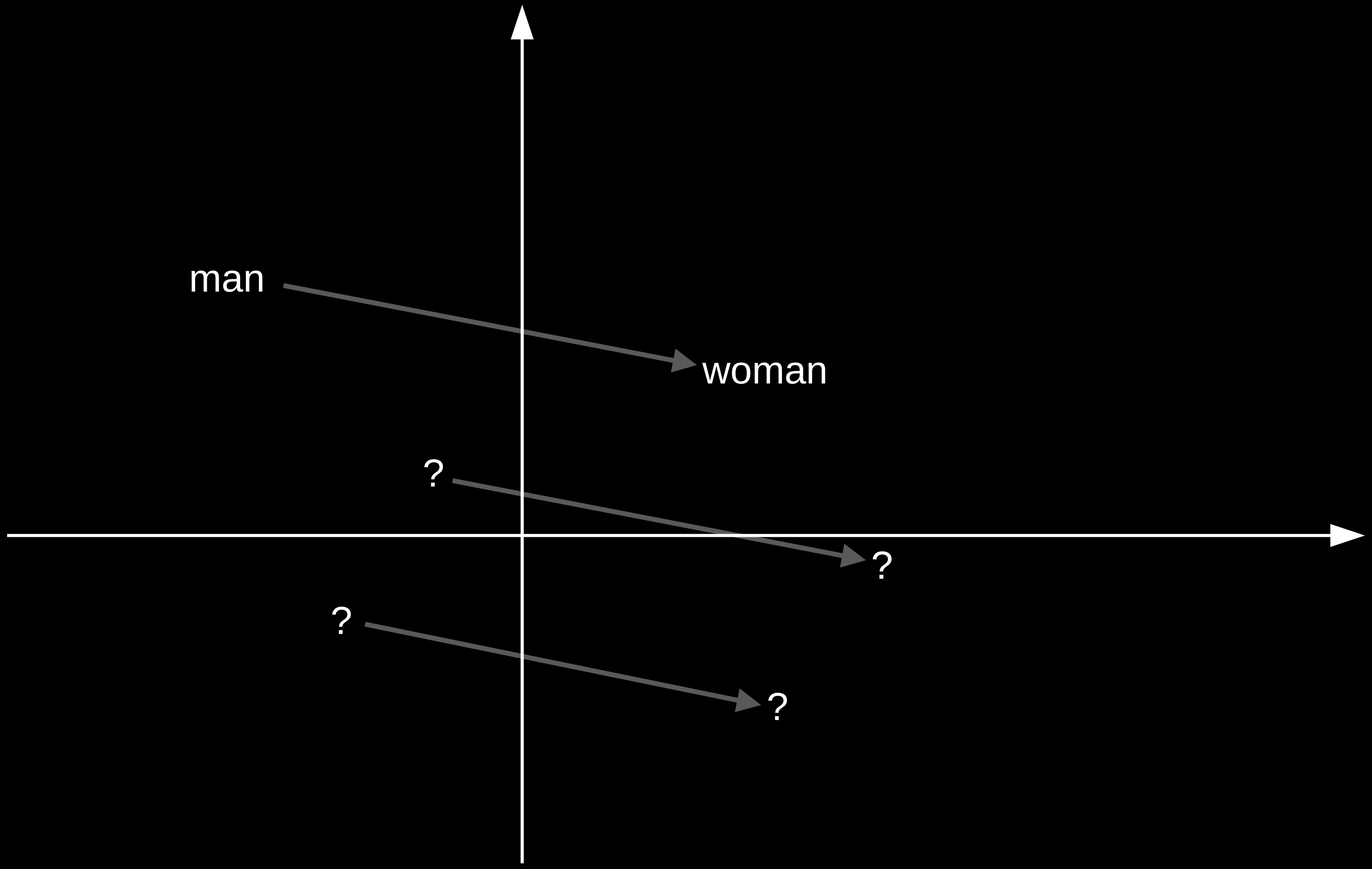
- **A Critique of Word Similarity as a Method for Evaluating Distributional Semantic Models** [pdf]
  Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds, David Weir



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

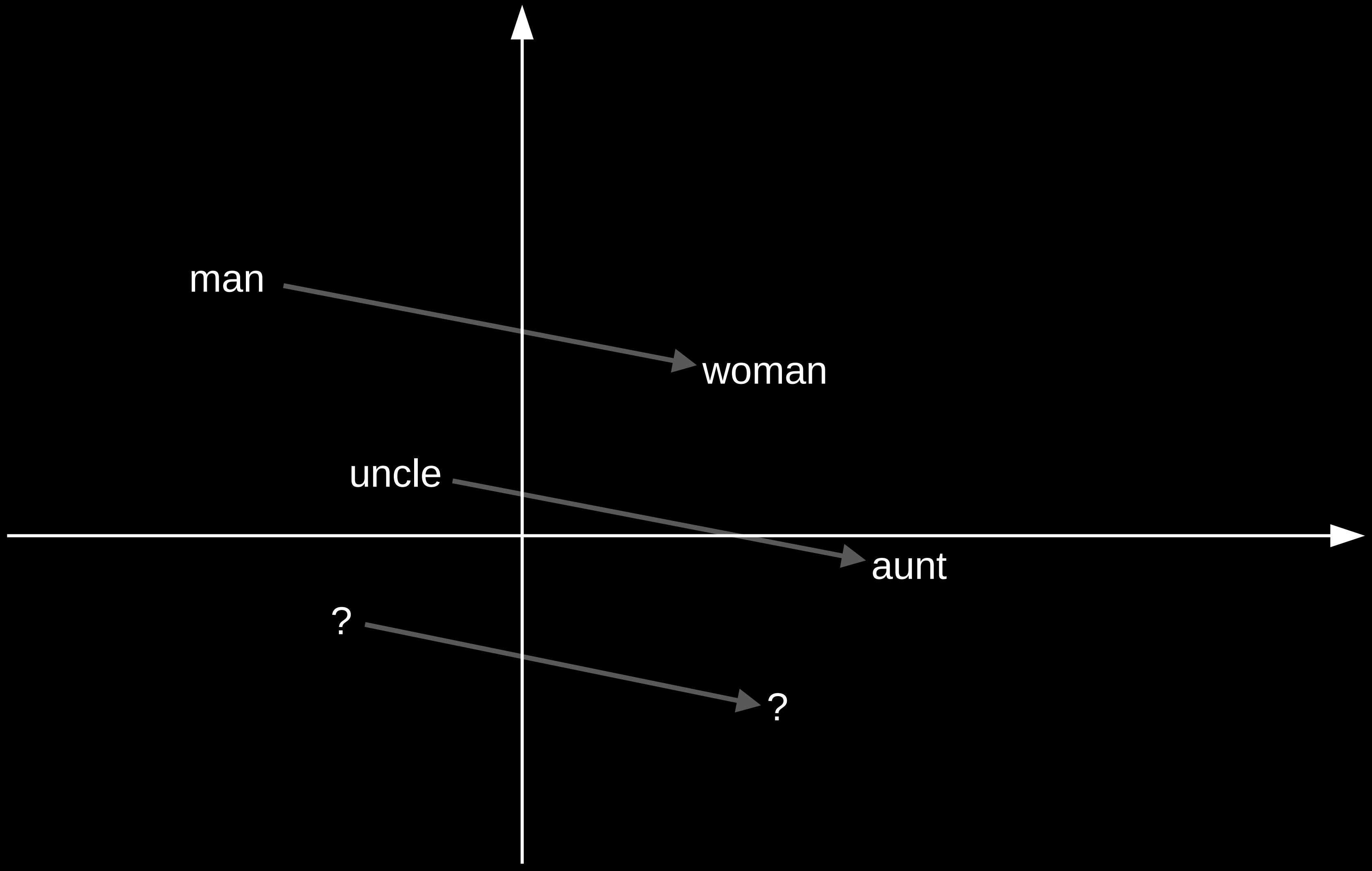
Word analogies revisited

- a:b:?:?



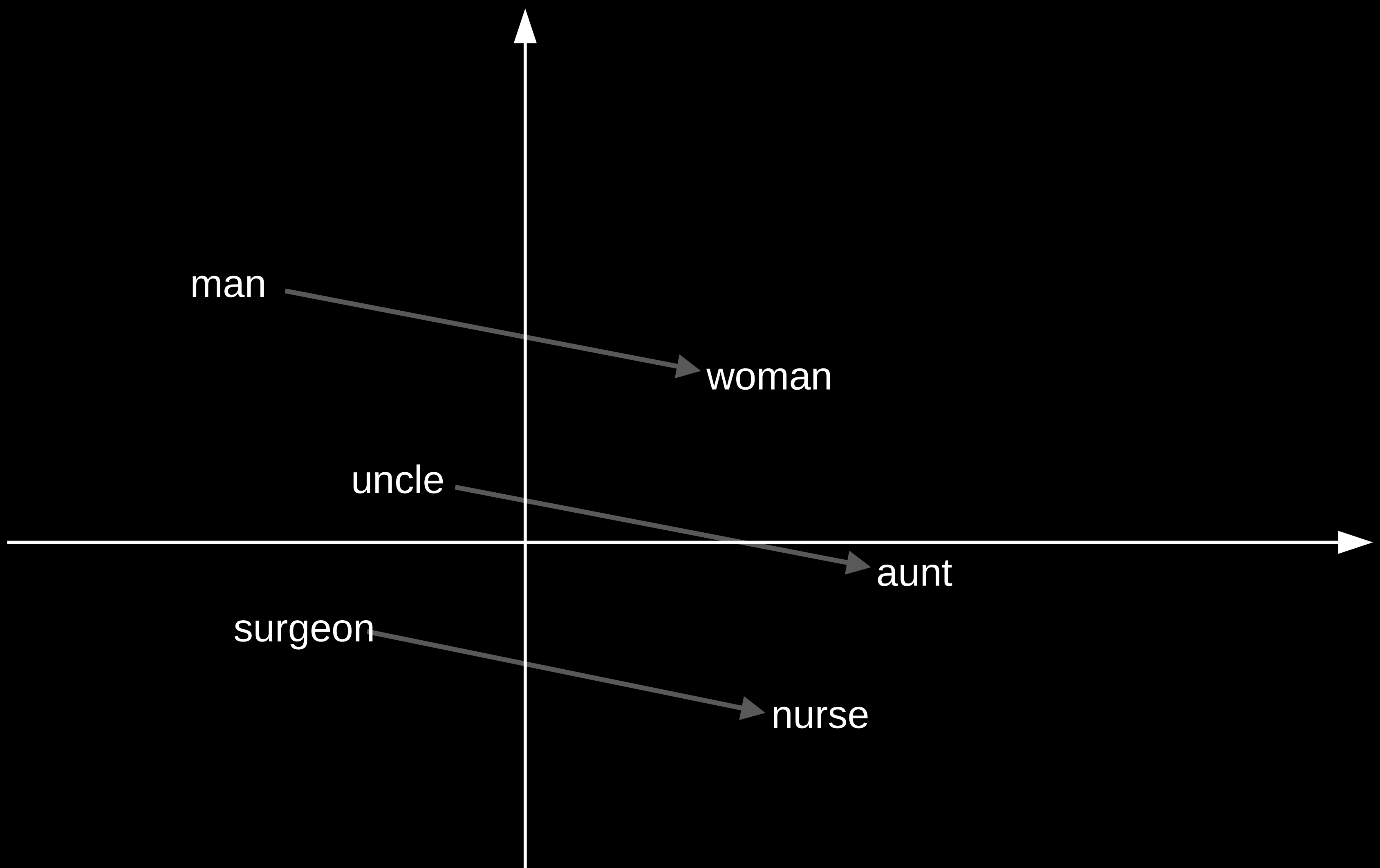
Word analogies revisited

- a:b:?:?



Word analogies revisited

- a:b:?:?



Limitations of word embeddings

- The quality of the embedding is determined by the corpus it is trained on

Limitations of word embeddings

- The quality of the embedding is determined by the corpus it is trained on
  - Potential recall issues (esp. for agglutinative languages)
    - Character level models
  - Polisemy (e.g. bank)
  - Multilinguality
  - Difficulties of evaluation
  - PCness (e.g. man : programmer :: women : X)
  - **Limited interpretability**

Sparse & continuous representations

apple [3.2 -1.5] → [0 1.7 0 0 -0.2 0]

...

banana [2.8 -1.6] → [0 1.1 0 0 -0.4 0]

...

door [-1.1 12.6] → [1.7 0 -2.1 0 0 -0.8]

…

zebra [0.8 0.5] → [0 0 1.3 0 -1.2 0]

Creating sparse word representations

- Assuming trained word embeddings w_i (i=1,…,|V|)

$$\min_{D \in C, A} \sum_{i=1}^{|V|} \| x_i - D a_i \|_2^2$$

Embedding vector ($\in \mathbb{R}^L$)

Dictionary ($\in \mathbb{R}^{LxK}$)

coefficients

Creating sparse word representations

- Assuming trained word embeddings w_i ($i=1,\ldots,|V|$)

$$\min_{D\in C,A} \sum_{i=1}^{|V|} \|x_i - D a_i\|_2^2 + \lambda \|a_i\|_1$$

Embedding vector ($\in \mathbb{R}^{L}$)

Dictionary ($\in \mathbb{R}^{L\mathrm{x}K}$)

Sparse coefficients

Sparsity inducing regularization

Creating sparse word representations

- Assuming trained word embeddings w_i (i=1,…,|V|)

$$\min_{D \in C, A} \sum_{i=1}^{|V|} \|x_i - D a_i\|_2^2 + \lambda \|a_i\|_1$$

Convex set of matrices s.t. $\forall \|d_i\| \leq 1$

Embedding vector ($\in \mathbb{R}^L$)

Dictionary ($\in \mathbb{R}^{LxK}$)

Sparse coefficients

Sparsity inducing regularization

Creating sparse word representations

- Assuming trained word embeddings w_i $(i=1,\dots,|V|)$

$$\min_{D\in C,A} \sum_{i=1}^{|V|} \|x_i - D a_i\|_2^2 + \lambda \|a_i\|_1$$

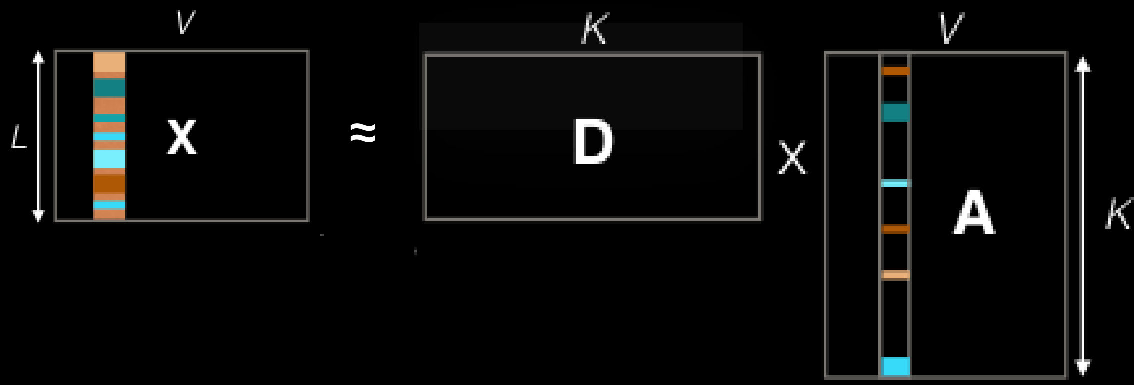
Convex set of matrices s.t. $\forall \|d_i\| \leq 1$

Embedding vector ($\in \mathbb{R}^L$)

Dictionary ($\in \mathbb{R}^{L\times K}$)

Sparse coefficients

Sparsity inducing regularization



“Classical” sequence labeling

- Calculate a set of (surface form) features using feature functions φ_j
  - φ_j could check for capitalization, suffixes, prefixes, neighboring words, etc.

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:						

“Classical” sequence labeling

- Calculate a set of (surface form) features using feature functions φ_j
  - φ_j could check for capitalization, suffixes, prefixes, neighboring words, etc.

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:	pre2=Fr suf2=it	pre2=fl suf2=es	pre2=li suf2=ke	pre2=a suf2=a	pre2=ba suf2=na	pre2=. suf2=.

“Classical” sequence labeling

- Calculate a set of (surface form) features using feature functions φ_j
  - φ_j could check for capitalization, suffixes, prefixes, neighboring words, etc.

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:	pre2=Fr suf2=it …	pre2=fl suf2=es …	pre2=li suf2=ke …	pre2=a suf2=a …	pre2=ba suf2=na …	pre2=. suf2=. ...

Sequence labeling using sparse word representation

- Rely on the sparse coefficients from α
  - $\phi(w_i) = \{ sign(\alpha_i[j])\, j \mid \alpha_i[j] \neq 0 \}$

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:						

Sequence labeling using sparse word representation

- Rely on the sparse coefficients from α
  - $\phi(w_i)=\{sign(\alpha_i[j])\,j \mid \alpha_i[j]\neq 0\}$
    - E.g. $\overrightarrow{Fruit}\approx 1.1\cdot\overrightarrow{d}_{28}-0.4\cdot\overrightarrow{d}_{171}$

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:						

Sequence labeling using sparse word representation

- Rely on the sparse coefficients from α
  - $\phi(w_i) = \{ sign(\alpha_i[j])\, j \mid \alpha_i[j] \neq 0 \}$
    - E.g. $\overrightarrow{Fruit} \approx 1.1 \cdot \overrightarrow{d}_{28} - 0.4 \cdot \overrightarrow{d}_{171}$

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:	P28 N171					

Sequence labeling using sparse word representation

- Rely on the sparse coefficients from α
  - $\phi(w_i) = \{ sign(\alpha_i[j])\, j \mid \alpha_i[j] \neq 0 \}$
    - E.g. $\overrightarrow{Fruit} \approx 1.1 \cdot \overrightarrow{d}_{28} - 0.4 \cdot \overrightarrow{d}_{171}$

X:	Fruit	flies	like	a	banana	.
Y:	NN	NN	VB	DT	NN	PUNCT
φ:	P28 N171	P77 P88 ...	N11 N62 ...	N88 N40 ...	P28 N210 ...	N21 P67 ...

Experimental setup

- Linear chain CRF (CRFsuite implementation)
- Part of Speech tagging
  - 12 languages from the CoNLL-X shared task
  - Google Universal Tag Set (12 tags)

Experimental setup

- Linear chain CRF (CRFsuite implementation)
- Part of Speech tagging
  - 12 languages from the CoNLL-X shared task
  - Google Universal Tag Set (12 tags)
- Hyperparameter settings
  - polyglot/w2v/Glove
  - m=64
  - k=1024
  - Varying λs

$$\min_{D \in C, \alpha} \sum_{i=1}^{|V|} \|w_i - D\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

Embedding vector ($\in \mathbb{R}^m$)

Dictionary ($\in \mathbb{R}^{m\text{x}k}$)

Sparse coefficients

Baselines

- Feature rich baseline (FR)
  - Standard feature set borrowed from CRFsuite
    - Previous, next word, word combinations, …
  - 2 variants:
    - Character+word level features (FR_{w+c})
    - Word level features alone (FR_w)

Baselines

- Feature rich baseline (FR)
  - Standard feature set borrowed from CRFsuite
    - Previous, next word, word combinations, …
  - 2 variants:
    - Character+word level features (FR_{w+c})
    - Word level features alone (FR_{w})

$FR_{w+c} \supset FR_{w}$

Baselines

- Feature rich baseline (FR)
  - Standard feature set borrowed from CRFsuite
    - Previous, next word, word combinations, …
  - 2 variants:
    - Character+word level features (FR_{w+c})
    - Word level features alone (FR_w)
- Brown clustering
  - Derive features from prefixes of Brown cluster IDs

Baselines

- Feature rich baseline (FR)
  - Standard feature set borrowed from CRFsuite
    - Previous, next word, word combinations, …
  - 2 variants:
    - Character+word level features (FR_{w+c})
    - Word level features alone (FR_w)
- Brown clustering
  - Derive features from prefixes of Brown cluster IDs
- Features from **dense** embeddings
  - $\phi(w_i) = \{ j : \alpha_i[j] \mid \forall j \in 1, \ldots, 64 \}$

Continuous vs. sparse embeddings

- Results averaged over 12 languages

	Dense	Sparse	
CBOW	88.30%	93.74%	
SG	86.89%	93.63%	

- Key inspections
  - polyglot > CBOW > SG > Glove

Continuous vs. sparse embeddings

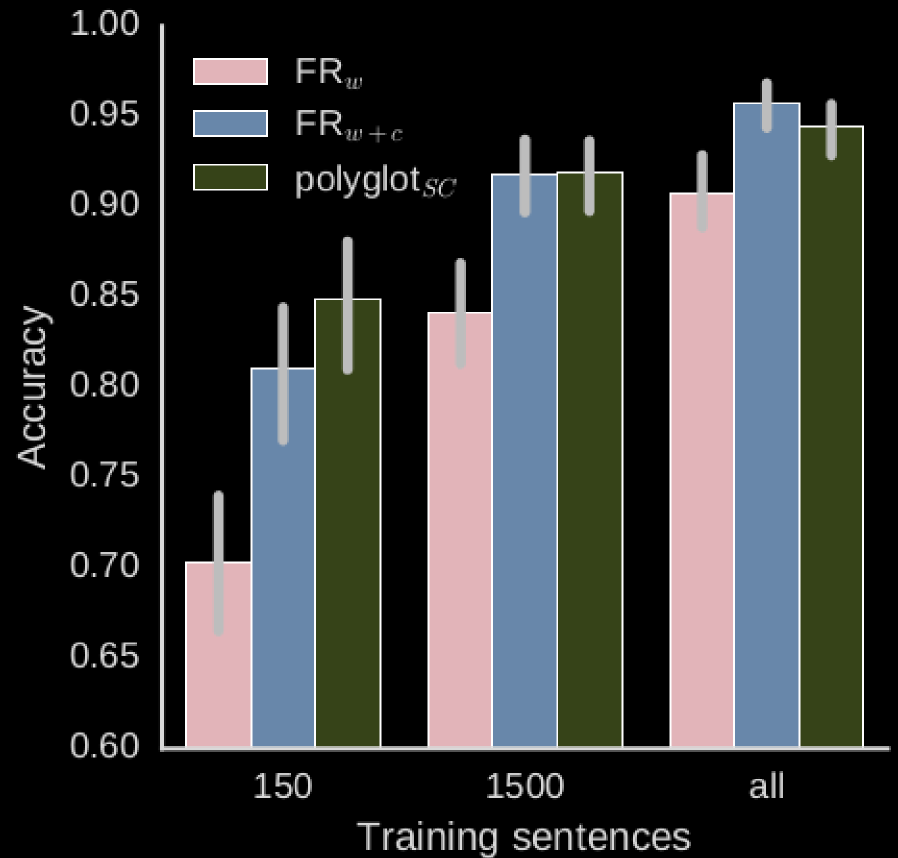
- Results averaged over 12 languages

	Dense	Sparse	Improvement
CBOW	88.30%	93.74%	+5.4
SG	86.89%	93.63%	+6.7

- Key inspections
  - polyglot > CBOW > SG > Glove
  - Sparse embeddings >> dense embeddings

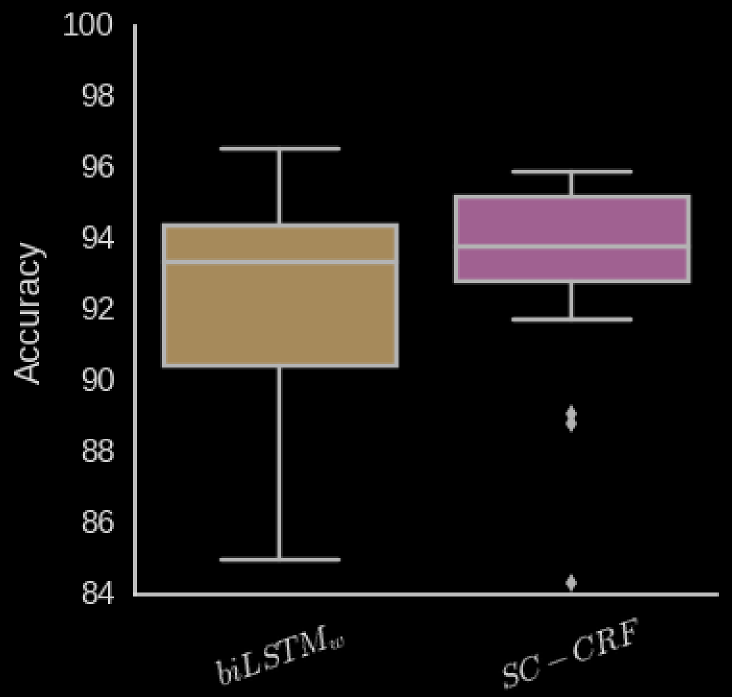
Experiments on generalization

- Training data artificially decreased
  - First 150 and 1500 sentences



Comparison with biLSTMs

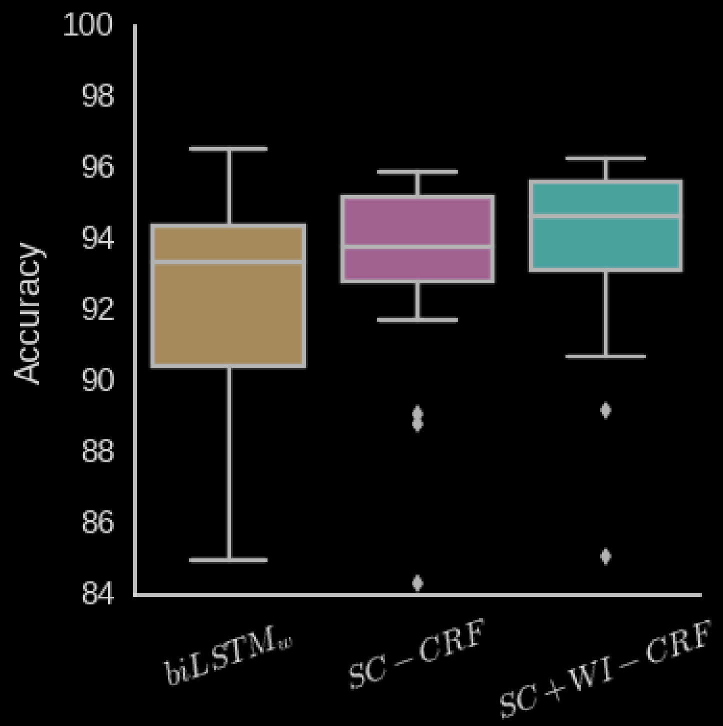
- POS tagging experiments on UD v1.2 treebanks
- Same settings as before (k=1024, λ=0.1)
- biLSTM results from *Plank et al. (2016)*



Method	Avg. accuracy
biLSTM_w	92.40%
SC-CRF	93.15%

Comparison with biLSTMs

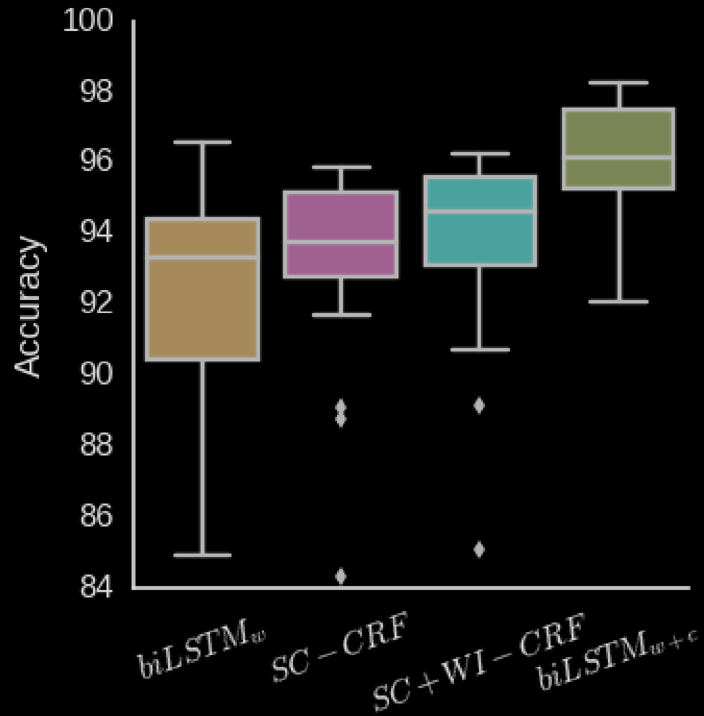
- POS tagging experiments on UD v1.2 treebanks
- Same settings as before (k=1024, λ=0.1)
- biLSTM results from *Plank et al. (2016)*



Method	Avg. accuracy
biLSTM_w	92.40%
SC-CRF	93.15%
SC+WI-CRF	93.73%

Comparison with biLSTMs

- POS tagging experiments on UD v1.2 treebanks
- Same settings as before (k=1024, λ=0.1)
- biLSTM results from *Plank et al. (2016)*



Method	Avg. accuracy
biLSTM_{w}	92.40%
SC-CRF	93.15%
SC+WI-CRF	93.73%
biLSTM_{w+c}	95.99%

Interpretability of sparse representation

- Does the discrete features (aka. the indices used in the reconstruction of a word) align with commonsense knowledge?
- Using the ConceptNet (CN) knowledge graph we quantified the extent to which certain discretized features co-occur with words of certain traits

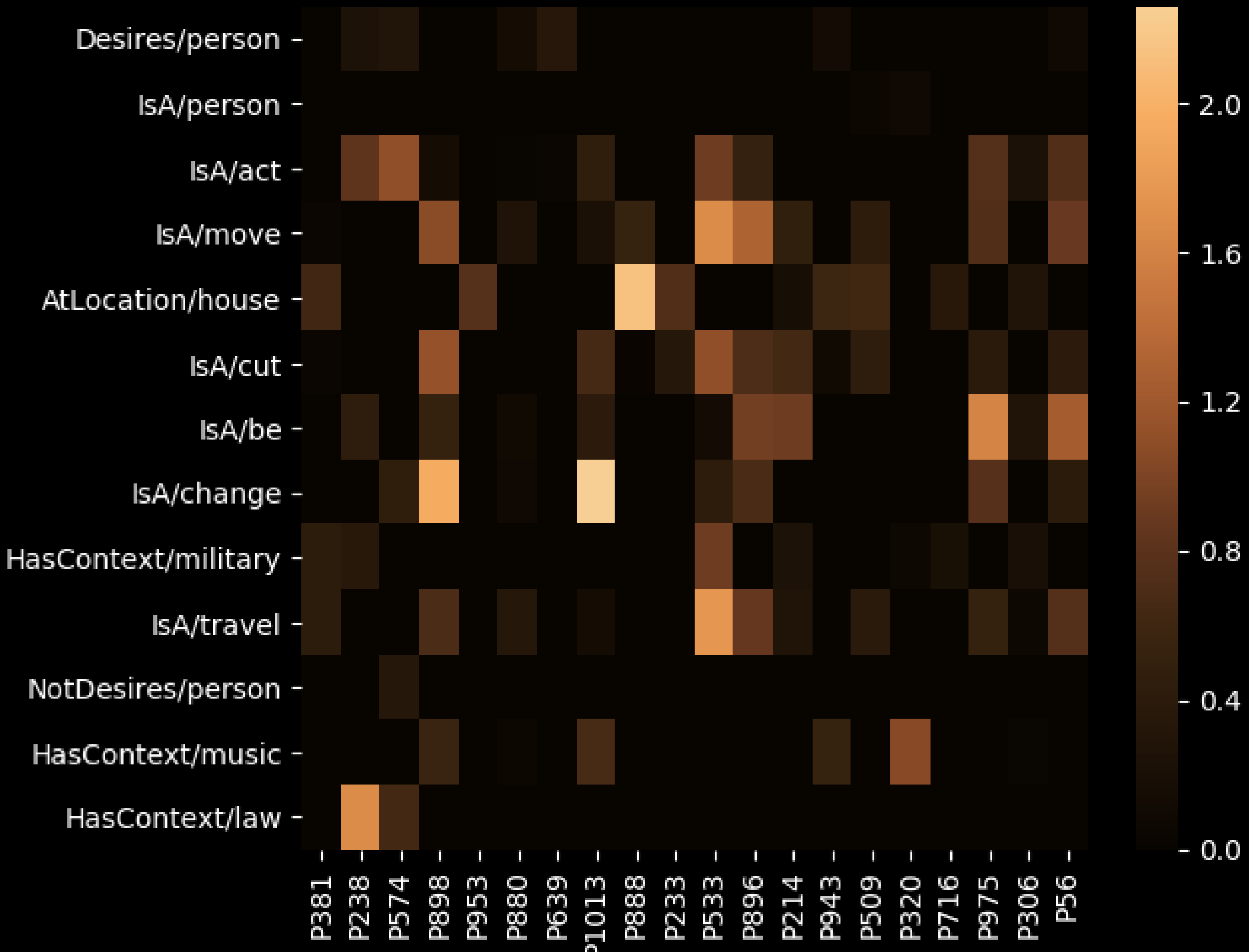
Interpretability of sparse representation

- Does the discrete features (aka. the indices used in the reconstruction of a word) align with commonsense knowledge?
- Using the ConceptNet (CN) knowledge graph we quantified the extent to which certain discretized features co-occur with words of certain traits
  - E.g. house → {4, **<u>31</u>**, **<u>91</u>**} and IsA('house', 'shelter') and hut → {7, **<u>31</u>**, 52, **<u>91</u>**} and IsA('hut', 'shelter')

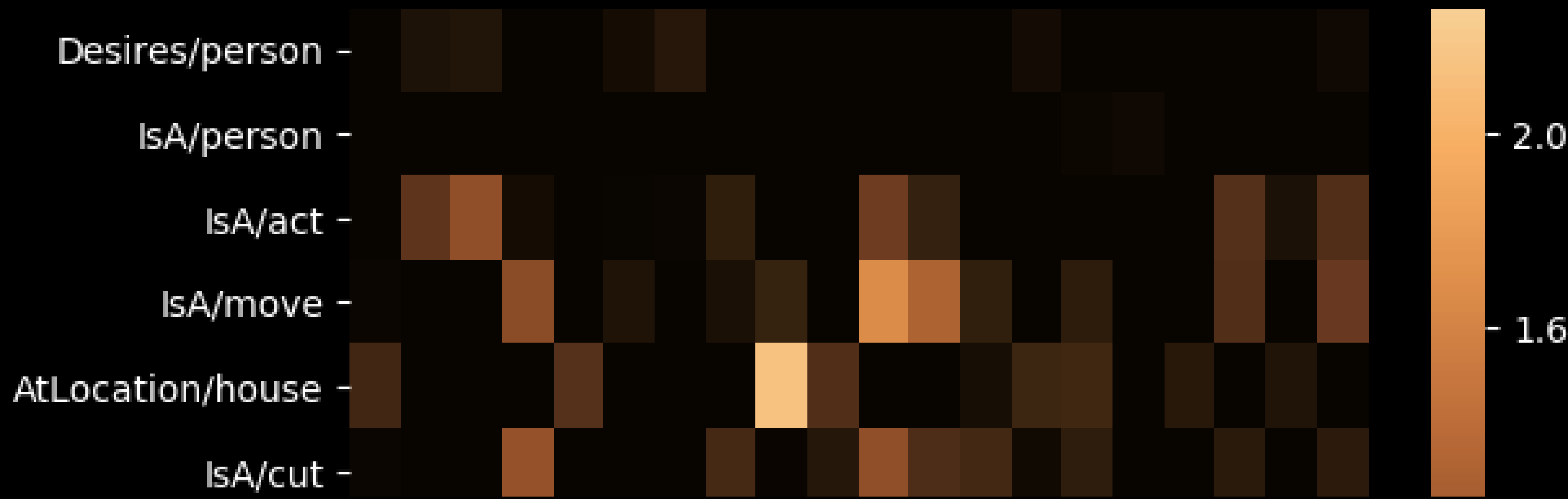
Interpretability of sparse representation

- Does the discrete features (aka. the indices used in the reconstruction of a word) align with commonsense knowledge?
- Using the ConceptNet (CN) knowledge graph we quantified the extent to which certain discretized features co-occur with words of certain traits
  - E.g. house → {4, **31**, **91**} and IsA('house', 'shelter') and hut → {7, **31**, 52, **91**} and IsA('hut', 'shelter')
    - The presence of base 31 and/or 91 seems to be a good indicator that something can be used as a shelter

Interpretability of sparse representation



Interpretability of sparse representation



Basis	Top-1	Top-2	Top-3	Top-4	Top-5	Most associated ConcepNet relation
P381	village	neighbourhood	neighborhood	fort	township	AtLocation/house
P238	amendment	decision	inquiry	obligation	petition	HasContext/law
P574	stability	coherence	sensitivity	separation	efficiency	IsA/act
P898	harden	darken	pierce	flatten	loosen	IsA/change
P953	coal	oil	food	cotton	grain	AtLocation/house

Conclusion

- Simple to implement, yet accurate analyzers
- Robustness across many languages (and tasks)
- Good generalization properties
- Encouraging results towards interpretability

Thank you for your attention!

berendg@inf.u-szeged.hu