Linguistic Regularities in Sparse and Explicit Word Representations

Omer Levy and Yoav Goldberg

R222 Presentation by Kaitlin Cunningham

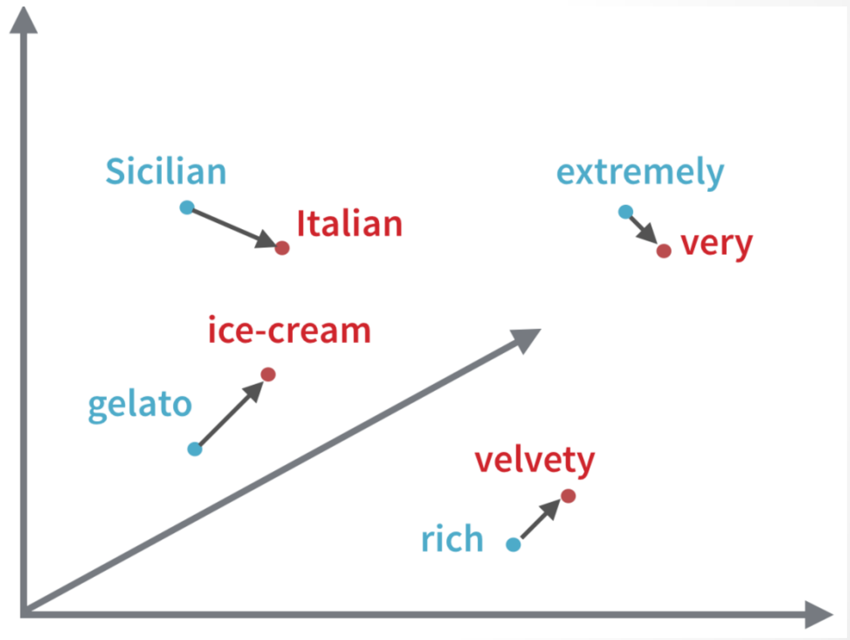
6 February 2017

Some theory

- Neural embeddings / word embeddings / distributed word representations:
  - Words represented as dense, real-valued vectors in $\mathbb{R}^d$
  - Embed an entire vocabulary into a low-dimensional linear space
  - Dimensions are latent continuous features

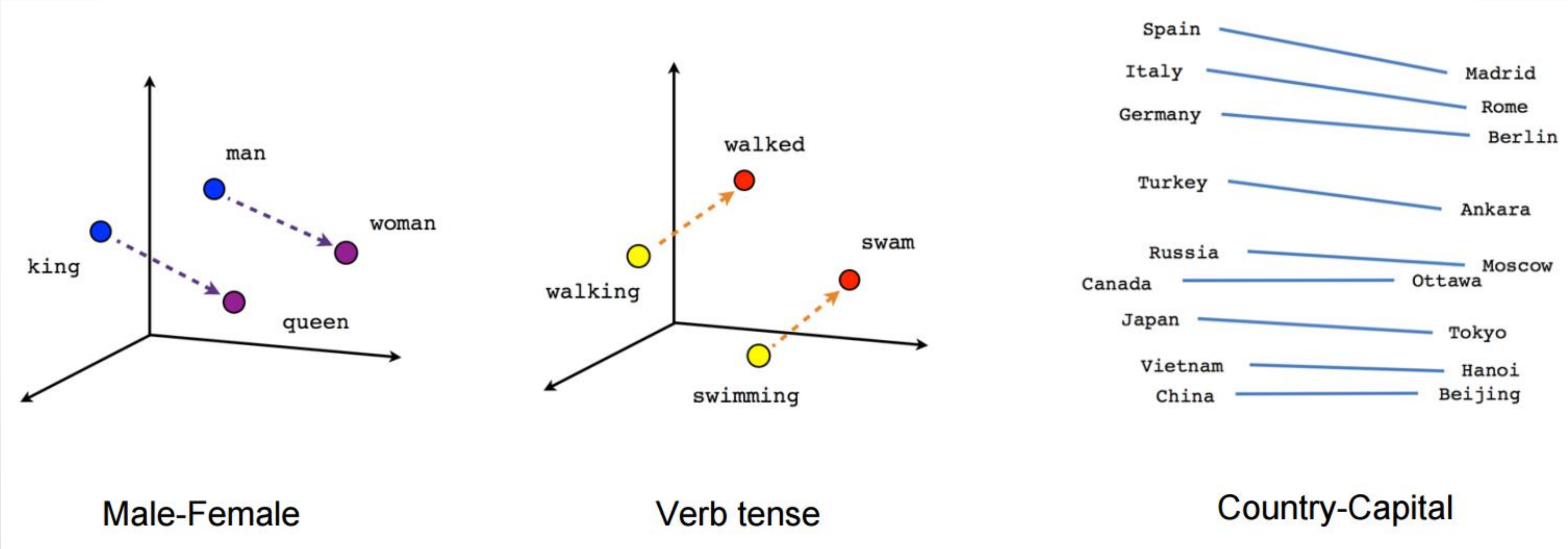
Some theory

- <u>Attributional similarities</u>:
  - words that appear in similar contexts will be close to each other in the vector space



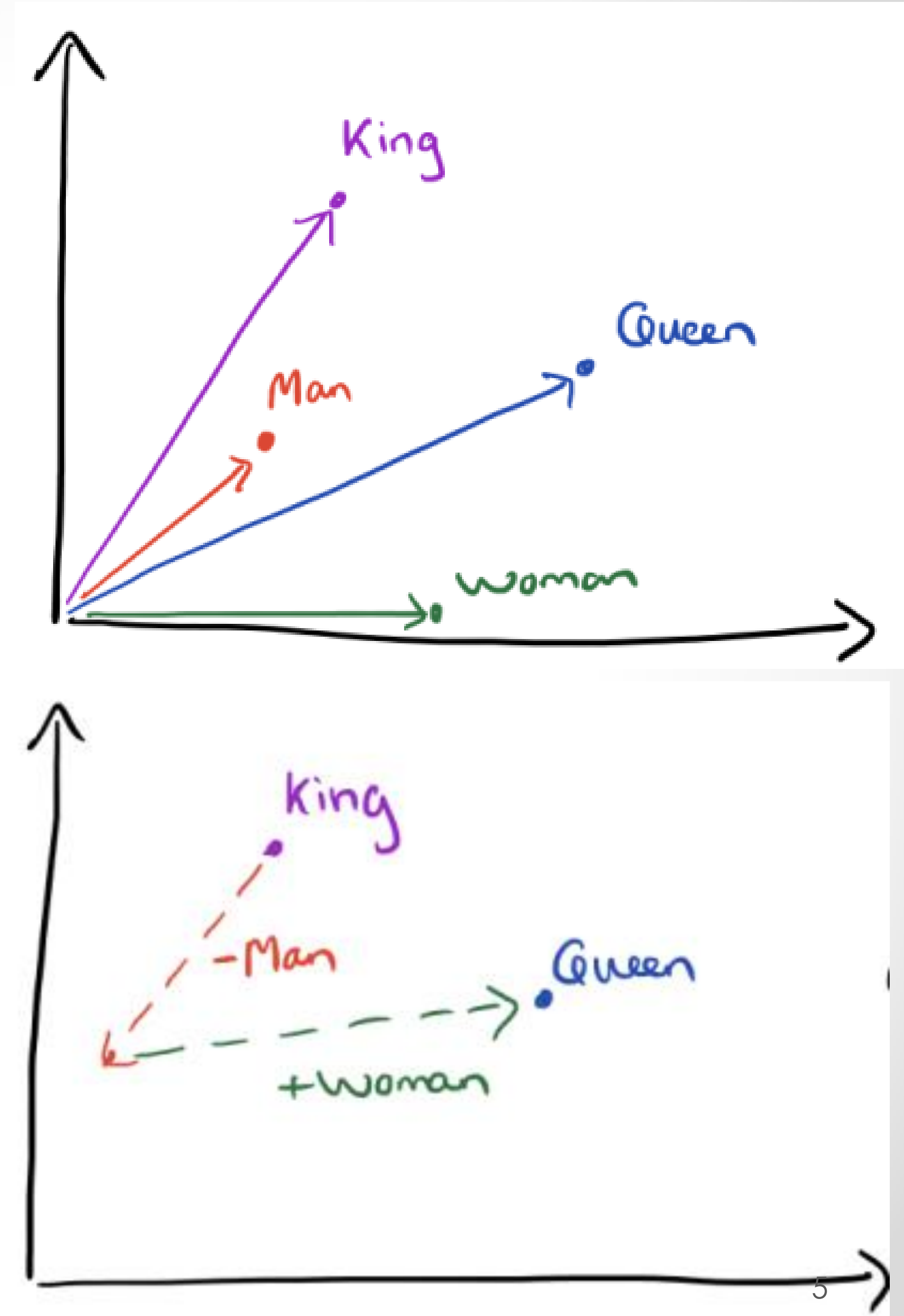
Some theory

- <u>Relational similarities</u>:
  - vectors can also encode linguistic relations like gender, tense



Some theory

- Relational similarities can be reflected in vector offsets between word pairs and accessed by using simple vector arithmetic:



Some theory

- An alternative to neural embeddings are explicit vector representations:
  - Each word is associated with a very high dimensional but sparse vector capturing the contexts in which it occurs
  - Each dimension corresponds to a context

This paper…

- Aims to show that the explicit vector space also encodes relational similarity information which can be recovered
- Contributes to the idea that the vector arithmetic method can be decomposed into a linear combination of three pairwise similarities
- And suggests a modified optimisation objective

Explicit vector space representations

→ For vocabulary *V* and a set of contexts *C*

The result is a $|V| \times |C|$ sparse matrix *S*

Where S_{ij} corresponds to the strength of the association between word *i* and context *j*

- The 'association' is measured by the positive pointwise mutual information (PPMI) metric

- The contexts are linear contexts which encompass the words surrounding the target word *w* within a window of 2 to each side

The analogy task

→ *a is to a* as b is to b**

- (1): 3CosAdd: $\arg\max_{b^* \in V} \left(\cos\left(b^*, b - a + a^*\right) \right)$

- (3): 3CosAdd:
$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b\right) - \cos\left(b^*, a\right) + \cos\left(b^*, a^*\right) \right)$$

- (2): PairDirection: $\arg\max_{b^* \in V} \left(\cos\left(b^* - b, a^* - a\right) \right)$

Basic setup

- English Wikipedia
- Word representations: WORD2VEC
- Evaluation datasets:
  - Open vocabularies: MSR, Google
  - Closed vocabulary: SEMEVAL

Micro-averaged accuracy

Macro-averaged accuracy

A reminder

- Derive explicit and neural embedded vector representations and compare their capacities to recover relational similarities using objectives 3CosAdd (3) and PairDirection (2)

Preliminary results

Representation	MSR	GOOGLE	SEMEVAL
Embedding	53.98%	62.70%	38.49%
Explicit	29.04%	45.05%	38.54%

Table 1: Performance of **3COSADD** on different tasks with the explicit and neural embedding representations.

Representation	MSR	GOOGLE	SEMEVAL
Embedding	9.26%	14.51%	44.77%
Explicit	0.66%	0.75%	45.19%

Table 2: Performance of **PAIRDIRECTION** on different tasks with the explicit and neural embedding representations.

Refining the maths

- (3): 3CosAdd:

$$\arg\max_{b^* \in V} \left(\cos\left(b^*, b\right) - \cos\left(b^*, a\right) + \cos\left(b^*, a^*\right) \right)$$

- (4): 3CosMul

$$\arg\max_{b^* \in V} \frac{\cos\left(b^*, b\right) \cos\left(b^*, a^*\right)}{\cos\left(b^*, a\right) + \varepsilon}$$

Results

Objective	Representation	MSR	GOOGLE	SemEval
3COSADD	Embedding	53.98%	62.70%	38.49%
	Explicit	29.04%	45.05%	38.54%
3COSMUL	Embedding	**59.09%**	66.72%	38.37%
	Explicit	56.83%	**68.24%**	38.67%

Table 3: Comparison of **3COSADD** and **3COSMUL**.

Error analysis

	Both Correct	Both Wrong	Embedding Correct	Explicit Correct
MSR	43.97%	28.06%	15.12%	12.85%
GOOGLE	57.12%	22.17%	9.59%	11.12%
ALL	53.58%	23.76%	11.08%	11.59%

71.9%

77.8%

Table 4: Agreement between the representations on open-vocabulary tasks.

Error analysis

	Relation	Embedding	Explicit
GOOGLE	capital-common-countries	90.51%	**99.41%**
	capital-world	77.61%	**92.73%**
	city-in-state	56.95%	**64.69%**
	currency	**14.55%**	10.53%
	family (gender inflections)	**76.48%**	60.08%
	gram1-adjective-to-adverb	**24.29%**	14.01%
	gram2-opposite	**37.07%**	28.94%
	gram3-comparative	**86.11%**	77.85%
	gram4-superlative	56.72%	**63.45%**
	gram5-present-participle	63.35%	**65.06%**
	gram6-nationality-adjective	89.37%	**90.56%**
	gram7-past-tense	**65.83%**	48.85%
	gram8-plural (nouns)	72.15%	**76.05%**
	gram9-plural-verbs	**71.15%**	55.75%
MSR	adjectives	45.88%	**56.46%**
	nouns	56.96%	**63.07%**
	verbs	**69.90%**	52.97%

Table 5: Breakdown of relational similarities in each representation by relation type, using 3CosMul.

Error analysis

Relation	Word	Emb	Exp
gram7-past-tense	who	0	138
city-in-state	fresno	82	24
gram6-nationality-adjective	slovak	39	39
gram6-nationality-adjective	argentine	37	39
gram6-nationality-adjective	belarusian	37	39
gram8-plural (nouns)	colour	36	35
gram3-comparative	higher	34	35
city-in-state	smith	1	61
gram7-past-tense	and	0	49
gram1-adjective-to-adverb	be	0	47
family (gender inflections)	daughter	8	47
city-in-state	illinois	3	40
currency	currency	5	40
gram1-adjective-to-adverb	and	0	39
gram7-past-tense	enhance	39	20

Table 6: Common default-behavior errors under both representations. Emb / Exp: the number of time the word was returned as an incorrect answer for the given relation under the embedded or explicit representation.

Why is all this cool?

- Words have several properties that affect how they relate to other words (i.e. their attributional similarities)
- Relational similarities are a composition of attributional similarities with each one reflecting a different aspect
    - Solving the analogy problem requires identifying the relevant aspects and changing one while preserving the other
- Explicit vector representations are just as good as neural embeddings, and less opaque

Inspecting the vectors

Aspect	Examples	Top Features
Female	$woman \odot queen$	estrid^{+1} ketevan^{+1} adeliza^{+1} nzinga^{+1} gunnhild^{+1} impregnate^{-2} hippolyta^{+1}
Royalty	$queen \odot king$	savang^{+1} uncrowned^{-1} pmare^{+1} sisowath^{+1} nzinga^{+1} tupou^{+1} uvea^{+2} majesty^{-1}
Currency	$yen \odot ruble$	devalue^{-2} banknote^{+1} denominated^{+1} billion^{-1} banknotes^{+1} pegged^{+2} coin^{+1}
Country	$germany \odot australia$	emigrates^{-2} 1943-45^{+2} pentathletes^{-2} emigrated^{-2} emigrate^{-2} hong-kong^{-1}
Capital	$berlin \odot canberra$	hotshots^{-1} embassy^{-2} 1925-26^{+2} consulate-general^{+2} meetups^{-2} nunciature^{-2}
Superlative	$sweetest \odot tallest$	freshest^{+2} asia's^{-1} cleveland's^{-2} smartest^{+1} world's^{-1} city's^{-1} america's^{-1}
Height	$taller \odot tallest$	regnans^{-2} skyscraper^{+1} skyscrapers^{+1} 6'4^{+2} windsor's^{-1} smokestacks^{+1} burj^{+2}

My two cents

The good

- Great contextualising paper

The less good

- Ignores syntactic relations
- Doesn't explain why/why not PPMI
- Ignores the non-improvement in SemEval 3CosMul results
- Very theoretical

?