Linguistic Regularities in Sparse and Explicit Word Representations

Omer Levy∗ and Yoav Goldberg

Presented by Haoyue Zhu

Previous work

- Word embedding : words -> $\mathbb{R}^d$
- Designed to capture *attributional similarities* between vocabulary items
- The effect is grouping of words that share semantic or syntactic properties

Mikolov et al. 2013

- capture the similarities between *pairs of words*
  - *Linguistic regularities / Relational similarities*
  - *e.g. gender relation, language-spoken-in relation, past-tense relation…*

    “man:woman”, “king:queen”; “france:french”,”china:chinese”; “go:went”,”play:played”

- Reflected in vector offsets between word pairs

$$\text{apples} - \text{apple} \approx \text{cars} - \text{car}$$

- Solve analogy questions of the form “a is to a* as b is to _”

$$\textit{queen} \approx \textit{king} - \textit{man} + \textit{woman}$$

The problem

- To what extent are the relational semantic properties a result of the **embedding** process?

- Alternative approach – *bag of context*
  - high dimensional but sparse vector
  - *Explicit* - each dimension directly corresponds to a particular context

Explicit Vector Space Representation

- |V| x |C| sparse matrix S
- S_{ij} : strength of the association between word *i* and context *j*
- PPMI metric

$$S_{ij} = PPMI(w_i, c_j)$$

$$PPMI(w,c) = \begin{cases} 0 & PMI(w,c) < 0 \\ PMI(w,c) & otherwise \end{cases}$$

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)} = \log \frac{freq(w,c)|corpus|}{freq(w)freq(c)}$$

Explicit Vector Space Representation

- Linear context
  - For sentence “a b c d e”
  - the contexts of the word c are a^{-2}, b^{-1}, d^{+1} and e^{+2}

- $|C| \approx 4|V|$

Vector Arithmetic

- 3COSADD

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

$$\cos\left(u, v\right) = \frac{u \cdot v}{||u|| \, ||v||}$$

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

- Reinterpreting

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

Vector Arithmetic

- PAIRDIRECTION

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

Empirical Setup

- Underlying Corpus and Preprocessing
  - English Wikipedia
  - Filtered non-alphanumeric tokens
  - Removed duplicates and sentence with less then 5 tokens
- Word Representation
  - Both embedding and explicit representation
  - 5 grams
  - Ignoring words < 100 times in corpus

Evaluation Datasets

- Open vocabulary – guess b* from entire vocabulary
  - ***MSR:*** 8000 analogy questions, morpho-syntactic relations categorized into adjectives, nouns and verbs

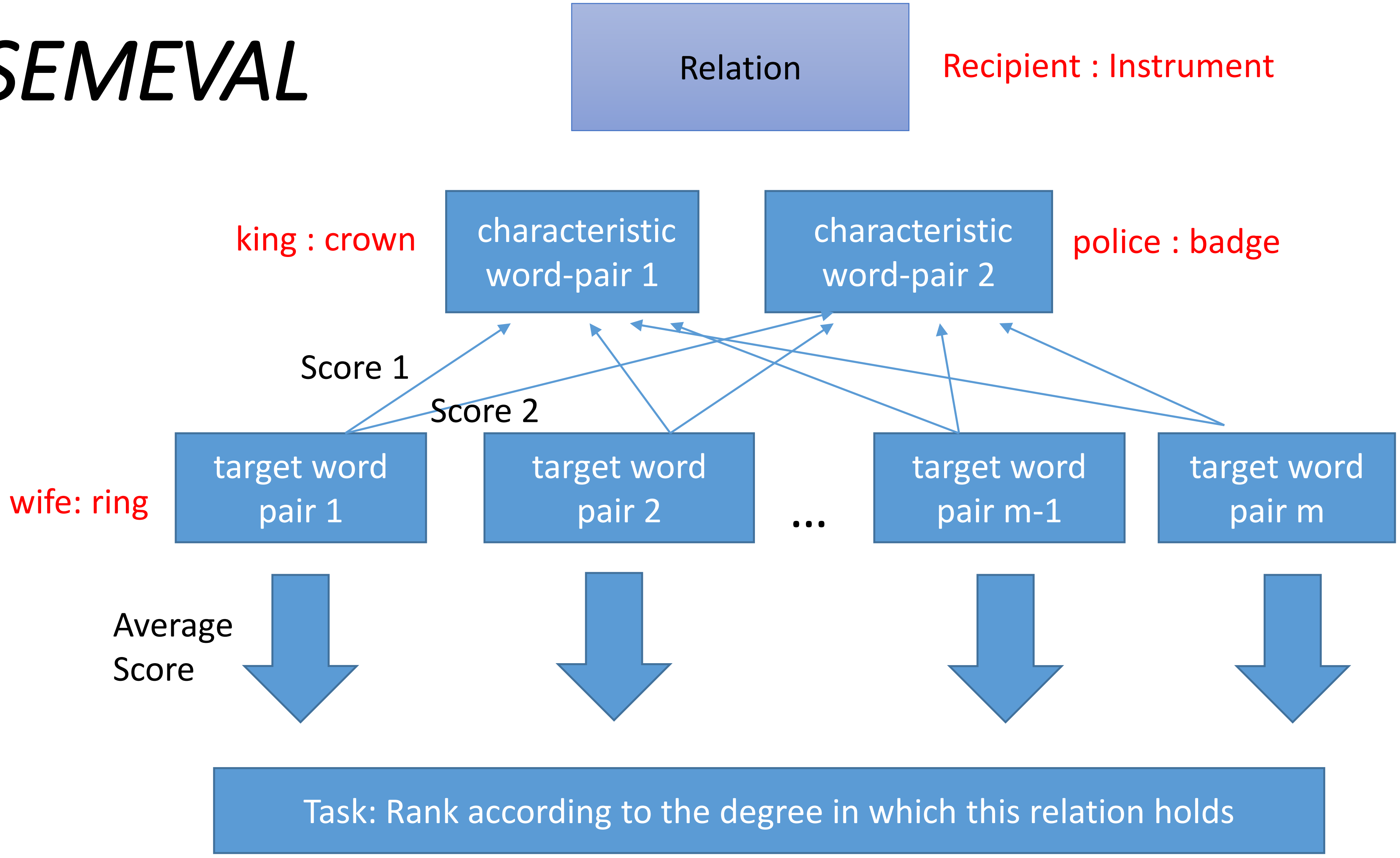
  - ***GOOGLE:*** 19544 questions, 7 semantic relations and 7 morpho-syntactic relations.

  \- Micro-averaged accuracy

- Closed vocabulary – ranking of candidate word pairs
  - ***SEMEVAL:*** 79 semantic relations

  \- Macro-averaged accuracy

SEMEVAL



Score 1 : “king is to crown as wife is to ring”
Score 2 : “police is to badge as wife is to ring”

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.

Scale Problem in 3COSADD

- Each term reflects a different aspect of similarity, and the different aspects have different scales.
  - “London is to England as Baghdad is to — ?”

$$\arg\max_{x \in V} \left(\cos(x, en) - \cos(x, lo) + \cos(x, ba)\right)$$

(EXP)	↑ England	↓ London	↑ Baghdad	Sum
Mosul	0.031	0.031	0.244	0.244
Iraq	0.049	0.038	0.206	0.217

(Explicit)

(EMB)	↑ England	↓ London	↑ Baghdad	Sum
Mosul	0.130	0.141	0.755	0.748
Iraq	0.153	0.130	0.631	0.655

(Embedding)

3COSMUL

- Switching from an additive to a multiplicative combination

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

($\varepsilon = 0.001$ is used to prevent division by zero)

Main Results

Objective	Representation	MSR	GOOGLE
3COSADD	Embedding	53.98%	62.70%
	Explicit	29.04%	45.05%
3COSMUL	Embedding	**59.09%**	66.72%
	Explicit	56.83%	**68.24%**

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

Error Analysis

- Agreement between Representations

	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%

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

If an answer is considered correct if it is correct in either representation, it can achieved an accuracy of 71.9% on the MSR dataset and 77.8% on GOOGLE.

Breakdown by Relation Type

	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.

Default-Behavior Errors

- one central representative word is provided as an answer to many questions of the same type
- Account for 49% of the errors in the explicit representation, and for 39% of the errors in the embedded representation
- Notable exceptions in explicit representation : “who”, “and”, “be” and “smith”
- 23.4% of the mistakes in past-tense relation are due to the explicit representation’s default answer of “who” or “and”, while 19% of the mistakes in the plural-verb relations are due to default answers of “is/and/that/who”.

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.

Verb-inflection Errors

- Requires recovering both
  - the correct inflection
  - the correct base word

- The morphological distinctions in verbs are much harder to capture than the semantics.

Interpreting Relational Similarities

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}

Table 7: The top features of each aspect, recovered by pointwise multiplication of words that share that aspect. The result of pointwise multiplication is an "aspect vector" in which the features common to both words, characterizing the relation, receive the highest scores. The feature scores (not shown) correspond to the weight the feature contributes to the cosine similarity between the vectors. The superscript marks the position of the feature relative to the target word.

Conclusion

- Similar to the neural embedding, the explicit vector also encodes a large amount of relational similarity which can be recovered in a similar fashion
- Neural embedding process is not discovering novel patterns, but rather is preserving the patterns
- The vector arithmetic method is mathematically equivalent to a linear combination of three pairwise similarities. It provides a better intuition on why we would expect the method to perform well on the analogy recovery task.
- It leads us to suggest a modified optimization objective, which outperforms the state-of-the-art at recovering relational similarities under both representations.

Thank you for listening

Questions?