Saint Petersburg State University

Estimating Syntagmatic Association Strength Using Distributional Word Representations

Ekaterina Protopopova, Grigoriy Bukia, Polina Panicheva, Olga Mitrofanova

DSMs and their usage

- classical:
  - LSA
  - HAL
  - …
- word embeddings:
  - word2vec
  - GloVe
  - …

- SemEval
  - semantic relatedness
  - compositional models
- RUSSE
  - relatedness
  - association

Estimating syntagmatic relations

герметичный **?** упаковка

вероятность **?** поломка

поднимать **?** тревога

- collocation strength – what about unseen pairs?
- compositionality

Method: count-based association

Confusion probability:

$$\mathrm{P}(w_1 \sim w_1) = \frac{\left|c(w_1) \cap c(w_2)\right|^2}{\left|c(w_1)\right|\left|c(w_2)\right|}$$

Basic association measure:

$$MI = \log \frac{p(w_1 w_2)}{p(w_1)p(w_2)}$$

Final measure:

$$Assoc(n,a) = \frac{\sum_{a_i \in c(n)} MI(n,a_i)\mathrm{P}(a,a_i) + \sum_{n_i \in c(a)} MI(n_i,a)\mathrm{P}(n,n_i)}{\sum_{a_i \in c(n)} \mathrm{P}(a,a_i) + \sum_{n_i \in c(a)} \mathrm{P}(n,n_i) + 1}$$

Method: word vector association

Skip-gram word embeddings

- Window size in [1, 10]
- N-dimensions in {25, 50, 100, 150, 300}

$$Comp(n,a) = \cos(n+a,n) = \cos(n,a)$$

Vecchi, E.M., Baroni, M., Zamparelli, R.: *(linear) maps of the impossible: capturing semantic anomalies in distributional space* (2011)

Kochmar, E., Briscoe, T.: *Capturing anomalies in the choice of content words in compositional distributional semantic space* (2013)

Method: word vector difference

What is the word (x) that is similar to *small* (x_c) in the same sense as *biggest* (x_b) is similar to *big* (x_a)?

y, most similar to x:

$$y = x_b - x_a + x_c$$

$$W2V_{rel}(a,n) = \max_{a_i, n_j \in K} \frac{< n, n_j - a_i + a >}{|n||n_j - a_i + a|}$$

Method: word vector difference

test combination	nearest difference combination
жевательная резинка – chewing gum	кавказский хребет – Caucasian chain цементная ступенька – cement step
копировальный центр — copy center	патрульный корабль – patrol ship
овощной салат – vegetable salad	конфетная коробка – a box of sweets гороховый суп – pea soup
чёрный кофе – black coffee	тёмное пиво – dark beer розовое мартини – pink martini

Pseudo-disambiguation test

Idea: 2 words can be associated if they co-occur in a small corpus of a general topic

Pekar, V.: *Distributivnaja model sochetaemostnyh ogranichenij glagolov [A distributional model of verbal selectional restrictions]* (2004)

Keyword	Correct pair	Incorrect pair
бдение ‘vigil’	ночной ‘nightly’	личный ‘personal’
ярость ‘fury’	немой ‘mute’	передний ‘front’
брюшной ‘abdominal’	пресс ‘muscle’	внешность ‘appearance’
шерстяной ‘woolen’	носочек ‘sock’	кража ‘robbery’

Pseudo-disambiguation test

1. Randomly select 500 **attributive NPs** for target nouns – (adj_1, noun)
2. **Delete** sentences containing them from the corpus
3. Select a **random** adj_2 for each pair:
   1) Not occurring with the target noun
   2) Having closest frequency to adj_1
4. (adj_1, noun) -> 'correct' pair
   (adj_2, noun) -> 'incorrect' pair
5. Same procedure for target adjectives

Training data

- Moshkov's library (lib.ru)

fiction loaded in **2009**

11.6M sentences - 140M tokens (**corpus A**)

fiction loaded in **2014**

345K sentences – 3.5M tokens(**B**)

- Testset annotated manually

Results

	W2Vrel	Comp	D
Acc	76%	81%	75%
Corr	88%	93%	84%

Acc – number of times adj_1 ranked higher than adj_2

Corr – *Acc* after manual annotation

Error analysis: common

Target	Reason	N	Examples
Noun	Occasional metaphor	45%	круглая сирота
			‘total orphan’
	General meaning	55%	европейский квартал
			‘european quarter’
Adjective	Occasional metaphor	45%	информационная чума
			‘informational boom’
	General meaning	55%	ограниченная сфера
			‘restricted sphere’

Error analysis: word vector difference

Target	Association frequency	N	Examples
Noun	high	25%	американская горка
	medium	20%	глухое недовольство
	low	55%	комсомольская прослойка
Adjective	high	27%	примерное поведение
	medium	20%	оливковый цвет
	low	53%	железнодорожный ключ

Error analysis: word vector association

Target	Association frequency	N	Examples
Noun	high	20%	последнее издыхание
	medium	25%	большая мышца
	low	55%	немой клон
Adjective	high	29%	трезвая голова
	medium	23%	копировальный центр
	low	48%	безработный фанатик

Errors: count-based measure

Target	Association frequency	N	Examples
Noun	high	26%	родовая схватка
	medium	20%	тёмное предчувствие
	low	54%	сухая конвульсия
Adjective	high	13%	жевательная резинка
	medium	50%	сумасшедшая история
	low	37%	суеверный закон

Conclusions

1. Very simple and not resource-greedy context count-based algorithm: 75-79%.
2. Word-embeddings perform better:75-82%.
3. Common errors are rare, but common error reasons are frequent:
   - Non-compositionality
   - Non-overlapping context of noun and adj
4. Word vector difference reveals regular patterns

Future work

1) Test set refinement
   - Manual correction of incorrect automatically obtained triplets
   - Test set extension
2) Algorithm refinement:
   - Accounting for metaphorical expressions
   - Differentiating between vague and specific meaning
   - Generalizing word-meanings into classes

Thank you!

Questions?