Evaluation of Word Embeddings

PA154 Language Modeling (8.2)

Pavel Rychlý

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

April 4, 2023

Word Embeddings

- many hyperparameters, diffrent training data
- different results even for same parameters and data
- what is better?
- how to compare quality of vectors?
- evaluate a direct outcome: word similarities

Sketch Engine Thesaurus

Lemma	Score	Freq
king	0.242	16,899
prince	0.213	6,355
charles	0.189	8,952
elizabeth	0.177	3,567
edward	0.176	6,484
mary	0.173	6,870
gentleman	0.171	6,274
lady	0.170	11,905
husband	0.167	11,669
sister	0.167	8,062
mother	0.164	27,536
princess	0.160	2,944
father	0.159	23,824
wife	0.157	18,308
brother	0.155	11,049
henry	0.151	6,699
daughter	0.150	11,216
anne	0.149	4,386

queen *(noun)*
British National Corpus (BNC) freq = 7,872 (70.10 per million)

Thesaurus evaluation

Gold standard

Source	Most similar words to *queen*
serelex	king, brooklyn, bowie, prime minister, mary, bronx, rolling stone, elton john, royal family, princess
Thesaurs.com	monarch, ruler, consort, empress, regent, female ruler, female sovereign, queen consort, queen dowager
SkE on BNC	king, prince, charles, elizabeth, edward, mary, gentleman, lady, husband, sister, mother, princess, father
SkE on enTenTen08	princess, prince, king, emperor, monarch, lord, lady, sister, lover, ruler, goddess, hero, mistress, warrior
word2vec on BNC	princess, prince, Princess, king, Diana, Queen, duke, palace, Buckingham, duchess, lady-in-waiting, Prince
powerthesaurus.org	empress, sovereign, monarch, ruler, czarina, queen consort, king, queen regnant, princess, rani, queen regent

Thesaurus evaluation

Gold standard

- very low inter-annotater agreement
- there are many directions of similarities
- existing gold standards not usable

Analogy queries

- evaluation of word embeddings (word2vec)
- ”a is to a^* as b is to b^*”, where b^* is hidden

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:
  Berlin –

Analogy queries

- evaluation of word embeddings (word2vec)
- ”*a* is to a^* as *b* is to b^*”, where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:
  Berlin – Germany

Analogy queries

- evaluation of word embeddings (word2vec)
- ”a is to a^* as b is to b^*”, where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:
  Berlin – Germany
  London –

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:
  Berlin – Germany
  London – England / Britain / UK ?

Analogy queries

- evaluation of word embeddings (word2vec)
- "a is to a^* as b is to b^*", where b^* is hidden
- syntactic: *good* is to *best* as *smart* is to *smarter*
- semantic: *Paris* is to *France* as *Tokyo* is to *Japan*
- agreement by humans:
  Berlin – Germany
  London – England / Britain / UK ?
- best match for linear combination of vectors:
  $\arg\max_{b^* \in V} cos(b^*, a^* - a + b)$

Analogy queries

Alternatives to cosine similarity

- $cos(x,y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$

Analogy queries

Alternatives to cosine similarity

- $cos(x, y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$
  $\arg\max_{b^* \in V}(cos(b^*, a^*) - cos(b^*, a) + cos(b^*, b))$
  (CosAdd)

Analogy queries
Alternatives to cosine similarity

- $cos(x, y) = \frac{v_x \cdot v_y}{\sqrt{v_x \cdot v_x}\sqrt{v_y \cdot v_y}}$
- $\arg\max_{b^* \in V} cos(b^*, a^* - a + b) =$
  $\arg\max_{b^* \in V}(cos(b^*, a^*) - cos(b^*, a) + cos(b^*, b))$
  (CosAdd)
- $\arg\max_{b^* \in V} \frac{cos(b^*, a^*) cos(b^*, b)}{cos(b^*, a)}$
  (CosMul)
- SkE uses Jaccard similarity instead of cosine similarity: JacAdd, JacMul

Thesaurus Evaluation

Results on capital-common-countries question set (462 queries)

	BNC		SkELL	
	count	percent	count	percent
CosAdd	58	12.6	183	39.6
CosMul	99	21.4	203	43.9
JacAdd	32	6.9	319	69.0
JacMul	57	12.3	443	95.9
word2vec	159	34.4	366	79.2

Results depends not only on data but also on the evaluation method.

Results on other corpora

More English corpora, using JacMul

Corpus	size (M)	correct
BNC	112	57
SkELL	1,520	443
araneum maius (LCL sketches)	1,200	224
enclueweb16	16,398	448
ententen 08	3,268	0
ententen 12	12,968	0
ententen 13	22,878	439

Problems of analogy queries

- Pair of words does not define an exact relation
- Berlin – Germany: capital, biggest city
- in what time?
- Canberra

Problems of analogy queries

- Pair of words does not define an exact relation
- Berlin – Germany: capital, biggest city
- in what time?
- Canberra, Rome

Problems of analogy queries

- Pair of words does not define an exact relation
- Berlin – Germany: capital, biggest city
- in what time?
- Canberra, Rome
- rare words/phrases
- Baltimore – Baltimore Sun: Cincinnati –

Problems of analogy queries

- Pair of words does not define an exact relation
- Berlin – Germany: capital, biggest city
- in what time?
- Canberra, Rome
- rare words/phrases
- Baltimore – Baltimore Sun: Cincinnati – Cincinnati Enquirer

Outlier detection

- list of words
- find the one which is not part of the cluster
- examples:
  - red, blue, green, dark, yellow, purple, pink, orange, brown

Outlier detection

- list of words
- find the one which is not part of the cluster
- examples:
  - red, blue, green, dark, yellow, purple, pink, orange, brown
  - t-shirt, sheet, dress, trousers, shorts, jumper, skirt, shirt, coat

Evaluating Outlier Detection

- original data set by Camacho-Collados, Navigli
- 8 pairs of 8 words in a cluster and 8 outliers
- $8 \times 8 = 64$ queries
- Accuracy – the percentage of successfully answered queries,
- Outlier Position Percentage (OPP) Score – average percentage of the right answer (Outlier Position) in the list of possible clusters ordered by their compactness

Problems of original data set

- English only
- needs extra knowledge
    - Mercedes Benz, BMW, Michelin, Audi, Opel, Volkswagen, Porsche, Alpina, Smart

Problems of original data set

- English only
- needs extra knowledge
  - Mercedes Benz, BMW, Michelin, Audi, Opel, Volkswagen, Porsche, Alpina, Smart
  - (Bridgestone, Boeing, Samsung, Michael Schumacher, Angela Merkel, Capri, pineapple)

Problems of original data set

- English only
- needs extra knowledge
  - Mercedes Benz, BMW, Michelin, Audi, Opel, Volkswagen, Porsche, Alpina, Smart
  - (Bridgestone, Boeing, Samsung, Michael Schumacher, Angela Merkel, Capri, pineapple)
  - Peter, Andrew, James, John, Thaddaeus, Bartholomew, Thomas, Noah, Matthew

Problems of original data set

- English only
- needs extra knowledge
  - Mercedes Benz, BMW, Michelin, Audi, Opel, Volkswagen, Porsche, Alpina, Smart
  - (Bridgestone, Boeing, Samsung, Michael Schumacher, Angela Merkel, Capri, pineapple)
  - Peter, Andrew, James, John, Thaddaeus, Bartholomew, Thomas, Noah, Matthew
  - January, March, May, July, Wednesday, September, November, February, June

Problems of original data set

- English only
- needs extra knowledge
    - Mercedes Benz, BMW, Michelin, Audi, Opel, Volkswagen, Porsche, Alpina, Smart
    - (Bridgestone, Boeing, Samsung, Michael Schumacher, Angela Merkel, Capri, pineapple)
    - Peter, Andrew, James, John, Thaddaeus, Bartholomew, Thomas, Noah, Matthew
    - January, March, May, July, Wednesday, September, November, February, June
    - tiger, dog, lion, cougar, jaguar, leopard, cheetah, wildcat, lynx
- mostly proper names (7 out of 8)

New data set: HAMOD

- 7 languages: Czech, Slovak, English, German, French, Italian, Estonian
- 128 clusters (8 words + 8 outliers)
- https://github.com/lexicalcomputing/hamod

New data set – example

Colors		Electronics	
Czech	English	Czech	English
červená	red	televize	television
modrá	blue	reproduktor	speaker
zelená	green	notebook	laptop
žlutá	yellow	tablet	tablet
fialová	purple	mp3 přehrávač	mp3 player
růžová	pink	mobil	phone
oranžová	orange	rádio	radio
hnědá	brown	playstation	playstation
dřevěná	wooden	blok	notebook
skleněná	glass	sešit	workbook
temná	dark	kniha	book
zářivá	bright	CD	CD
pruhovaný	striped	energie	energy
puntíkovaný	dotted	světlo	light
smutná	sad	papír	paper
nízká	low	ráno	morning

Evaluation

- 9 clusters only, 72 queries

	OOP	Accuracy
Czes2	92.2	70.8
czTenTen12	93.4	79.2
csTenTen17	94.3	81.9
czTenTen12 (fasttext)	97.7	87.5
Czech Common Crawl	98.1	95.8

Construction

- each human evaluator goes through all the sets (only once) for their native language
- 1 exercise: 8 inliers + 1 outlier (randomly chosen from the list of outliers for each set)
- in each turn, the evaluator selects the outlier
- simple web interface for the exercise
- Inter-Annotator Agreement: Estonian 0.93, Czech 0.97