Vektorová reprezentácia slov využívajúca morfológiu

Mária Vajdová
Vedúci: RNDr. Michal Forišek, PhD.
Konzultant: Mgr. Vladimír Boža

Fakulta matematiky, fyziky a informatiky, UK

21. júna 2017

Prehľad

Jazykové pojmy

- morféma - základná významová jednotka v slove (základ slova, prípona. . .)
- ngram (znakový) - postupnosť n za sebou idúcich znakov v slove

Strojová reprezentácia slov

reprezentácia slova pomocou číselného vektora

motivácia:

- podobnosť slov
- podobnosť textov
- predspracovanie pre iné algoritmy - porozumenie významu textu, poznávanie hlasu, strojový preklad

Jednoduché reprezentácie

- vektor dĺžky slovníka so samými nulami okrem 1 na poradovom mieste slova (veľkosť vektora slova = veľkosť slovníka)
- skrytá sémantická analýza - reprezentácia dokumentov pomocou batohu slov, následné zníženie dimenzií
- náhodné indexovanie - náhodné vektory pre všetky slová, zosumovanie s okolitými slovami

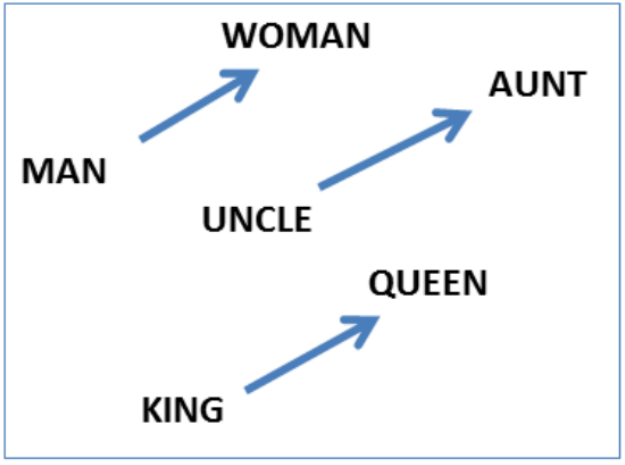
Word2Vec

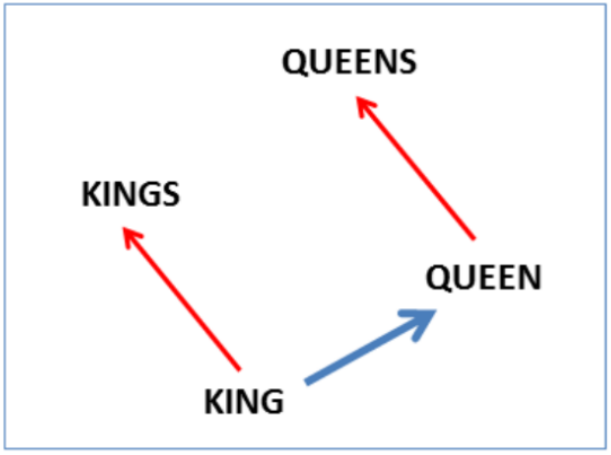
- vektor reálnych čísel ľubovoľnej dĺžky (zvyčajne 100)
- kosínusová podobnosť vektorov $\sim$ podobnosť dvoch slov
- ak sa dve slová vyskytujú často v podobnom kontexte sú podobné:
  - *I eat **fruit** salad. I eat **vegetable** salad.*
  - *I grow **fruit** in garden. I grow **vegetable** in garden.*[1]
- nevýhoda - pre slovo, ktoré sa nenachádzalo v trénovacej množine nevieme určiť vektor

$$\textit{woman} \rightarrow [-0.7231, 1.8392, 0.6352, \ldots, -1.17384, 2.3491, -1.2647]$$

Word2Vec - aritmetika

$$v(king) - v(man) + v(woman) \sim v(queen)$$



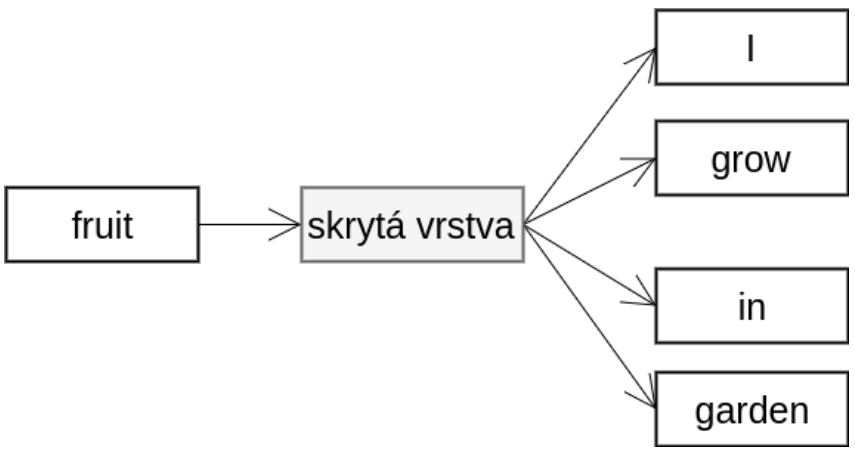


[2]

Word2Vec - architektúra siete

- neurónová sieť, ktorá sa snaží na základe slova predpovedať okolité slová (Skip-gram)
- tréning na čistom texte v danom jazyku, napr. Wikipédia dump

*I grow **fruit** in garden.*



Word2Vec - architektúra siete

snažíme sa maximalizovať priemerný log. pravdepodobnosti:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t),$$

$$p(w_O|w_I) = \frac{e^{{v_{wO}}'^{\top} v_{wI}}}{\sum_{w=1}^{W} e^{v'_w v_{wI}}},$$

vylepšenia - negatívne vzorkovanie, hierarchický softmax

Podobné práce

reprezentácia slova:

- slovný vektor + vektory morfém
  - učenie morfém s učiteľom, príp. slovník morfém [3]
  - učenie morfém bez učiteľa [4]
- slovný vektor + vektory ngramov [5]
- rekurentné znakové siete

možnosť reprezentovať nové slová, ALE veľká pamäťová náročnosť:
vektory pre slová + vektory pre ngramy/morfémy

Cieľ práce

vytvoriť reprezentáciu slov, ktorá:

- má nízku pamäťovú náročnosť (malý počet vektorov)
- ponúka dostatočne dobrú reprezentáciu (čo najviac sa priblížiť k existujúcim reprezentáciam)
- je schopná predpovedať neznáme slová
- nájsť čo najoptimálnejšie riešenie spĺňajúce dané vlastnosti

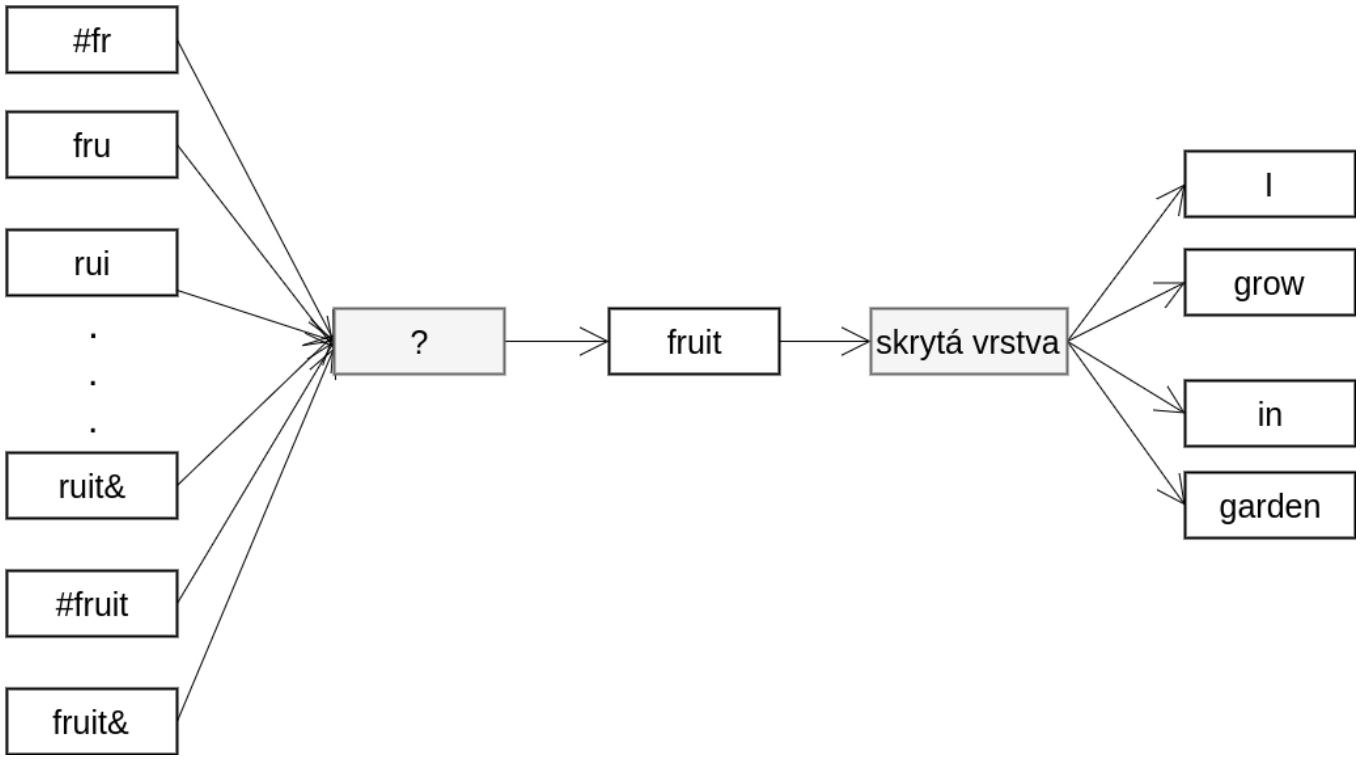
takýto model slovných vektorov s menším počtom parametrov by mohol umožňovať následné rýchlejšie trénovanie zložitejších modelov, kde by slúžil ako vstup

Náš prístup

- vytvorenie vektorov iba pre ngramy
- jednoduché delenie slova (netreba sa učiť morfémy)

woman $\rightarrow$ #wo, wom, oma, man, an&, #wom, woma, oman, man&, #woma, woman, oman&,#woman, woman&

Architektúra siete



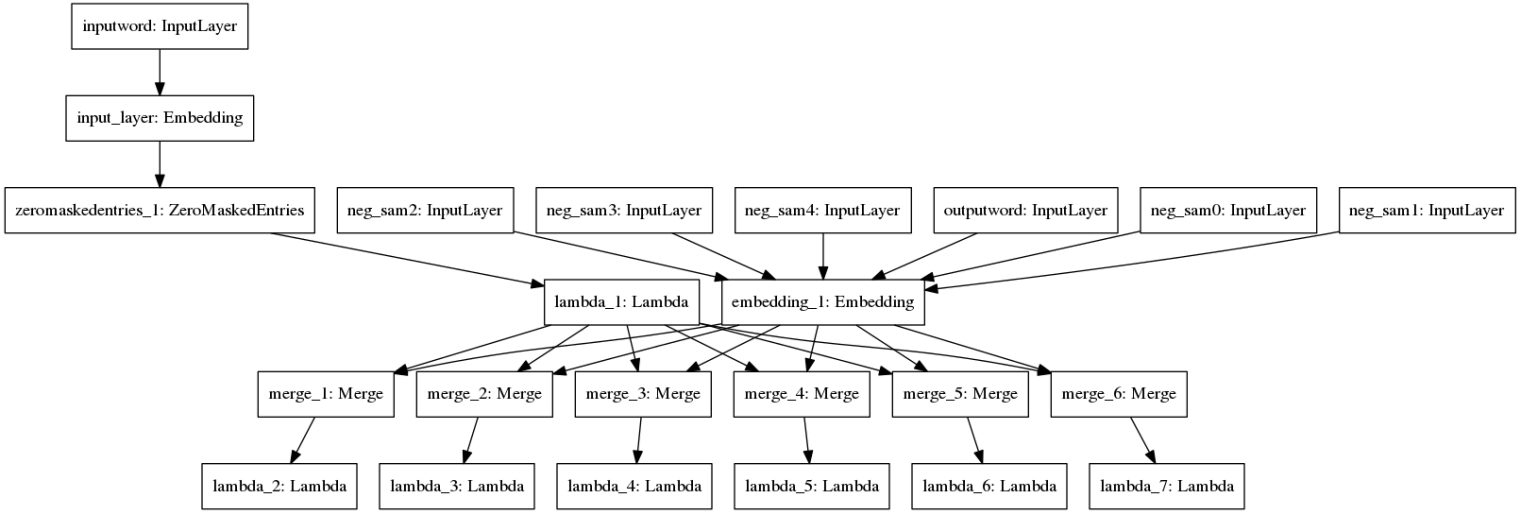
Ako spájať ngramy?

Spájanie ngramov

- suma
- max,min zreťazenie
- konvolučná sieť - v akom poradí vkladať slová?
  1. ngramy dĺžky 3, ngramy dĺžky 4, ngramy dĺžky 5, ngramy dĺžky 6, nulové vektory
  2. ngramy dĺžky 3, nulové vektory, ngramy dĺžky 4, nulové vektory ...
  3. 4 samostatné konvolúcie
- rekurentné siete (LSTM, GRU) - rovnaký problém ako pri konv.

Detaily implementácie

- programovací jazyk Python
- knižnica Keras
- 50 000 ngramov, dĺžka ngramov 3-6
- dĺžka slovných vektorov 100
- trénované na anglickej wikipédii



Výsledky - malý dataset

Názov triedy modelu	EN-WS353 20 mil.	EN-RW 20 mil.
WordVectorModel (Word2Vec)	*39.8*	*9.8*
NgramSumModel	36.4	27.9
NgramConvModel	**39.0**	**28.5**
NgramConvGatedModel	32.9	26.5

Tabuľka: Výsledky jednotlivých modelov

EN-WS353 - dataset podobnosti bežných anglických slov
EN-RW dataset podobností zriedkavých slov

Výsledky - veľké datasety

Názov triedy modelu	EN-WS353 50 mil.	EN-RW 50 mil.	EN-WS353 100 mil.	EN-RW 100 mil.
WordVectorModel (Word2Vec)	*56.1*	*16.5*	*62.9*	*39.0*
NgramSumModel	38.5	30.5	42.4	31.2
NgramConvModel	**40.7**	**29.9**	**52.2**	**31.5**
NgramConvGatedModel	37.0	28.6	51.0	31.5

Tabuľka: Výsledky úspešnejších modelov

EN-WS353 - dataset podobnosti bežných anglických slov
EN-RW dataset podobností zriedkavých slov

Hľadanie dôležitých ngramov

dôvody hľadania dôležitých ngramov:

- lepšie porozumenie modelu
- model založený na menšom počte ngramov $\rightarrow$ ešte rýchlejšie trénovanie

použité metódy:

- LIME (vysvetlenie predikcii klasifikátora) - aproximujeme predikčnú funkciu jednoduchšou funkciou [5]
- substitúcia ngramov - substituujeme jendotlivé ngramy v slove za náhodne slová a porovnávame vzniknuté vektory

model založený na najdôležitejších ngramoch

- pomocou oboch algoritmov sme vyextrahovali 10000 najdôležitejších ngramov
- vytvorili sme modely, ktoré využívali iba tento obmedzený počet ngramov

Výsledky

percentuálna podobnosť medzi našimi výsledkami a výsledkami datasetov podľa Pearsonovho kor. keoficientu

Názov triedy modelu	Počet parametrov modelu	EN-WS353 20 mil.	EN-RW 20 mil.
NgramConvModel	5 921 700	39.0	28.5
LIME model	1 740 100	23.5	17.1
Substitučný	1 740 100	36.3	22.2

Tabuľka: Výsledky LIME a substitučného modelu

EN-WS353 - dataset podobnosti bežných anglických slov
EN-RW dataset podobností zriedkavých slov

Záver

- natrénovaný model na 5 921 700 parametroch, oproti 9 144 900 bežných
- pri 10 000 ngramoch iba 1 740 100 parametrov
- možnosti ďalšieho využitia - extrakcia morf. informácie, stemmer a pod.

References


Mikolov, Tomas and Chen, Kai and Corrado, Greg and Dean, Jeffrey (2013)
Efficient estimation of word representations in vector space
arXiv preprint arXiv:1301.3781


Mikolov, Tomas and Yih, Wen-tau and Zweig, Geoffrey (2013)
Linguistic Regularities in Continuous Space Word Representations.
Hlt-naacl, vol. 13, p746–751


Luong, Thang and Socher, Richard and Manning, Christopher D (2013)
Better word representations with recursive neural networks for morphology.
CoNLL, p104–113


Soricut, Radu and Och, Franz Josef (2015)
Unsupervised Morphology Induction Using Word Embeddings.
HLT-NAACL, vol. 13, p1627–1637


Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas (2016)
Enriching word vectors with subword information
arXiv preprint arXiv:1607.04606


Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos (2016)
Why Should I Trust You?: Explaining the Predictions of Any Classifier
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,1135–1144

Ďakujem za pozornosť.

???