Improving Word Representations via Global Context and Multiple Word Prototypes

Eric H. Huang

2012

a new architecture

- incorporating both local and global document context
- accounts for homonymy and polysemy by learning multiple embeddings per word

Minimizing the ranking loss

$$C_{s,d} = \sum_{\omega \in V} \max(0, 1 - g(s,d) + g(s^{\omega}, d))$$

s: a word sequence

d: document in which s occurs

g(): scoring function

s^w: s with the last word replaced by word w

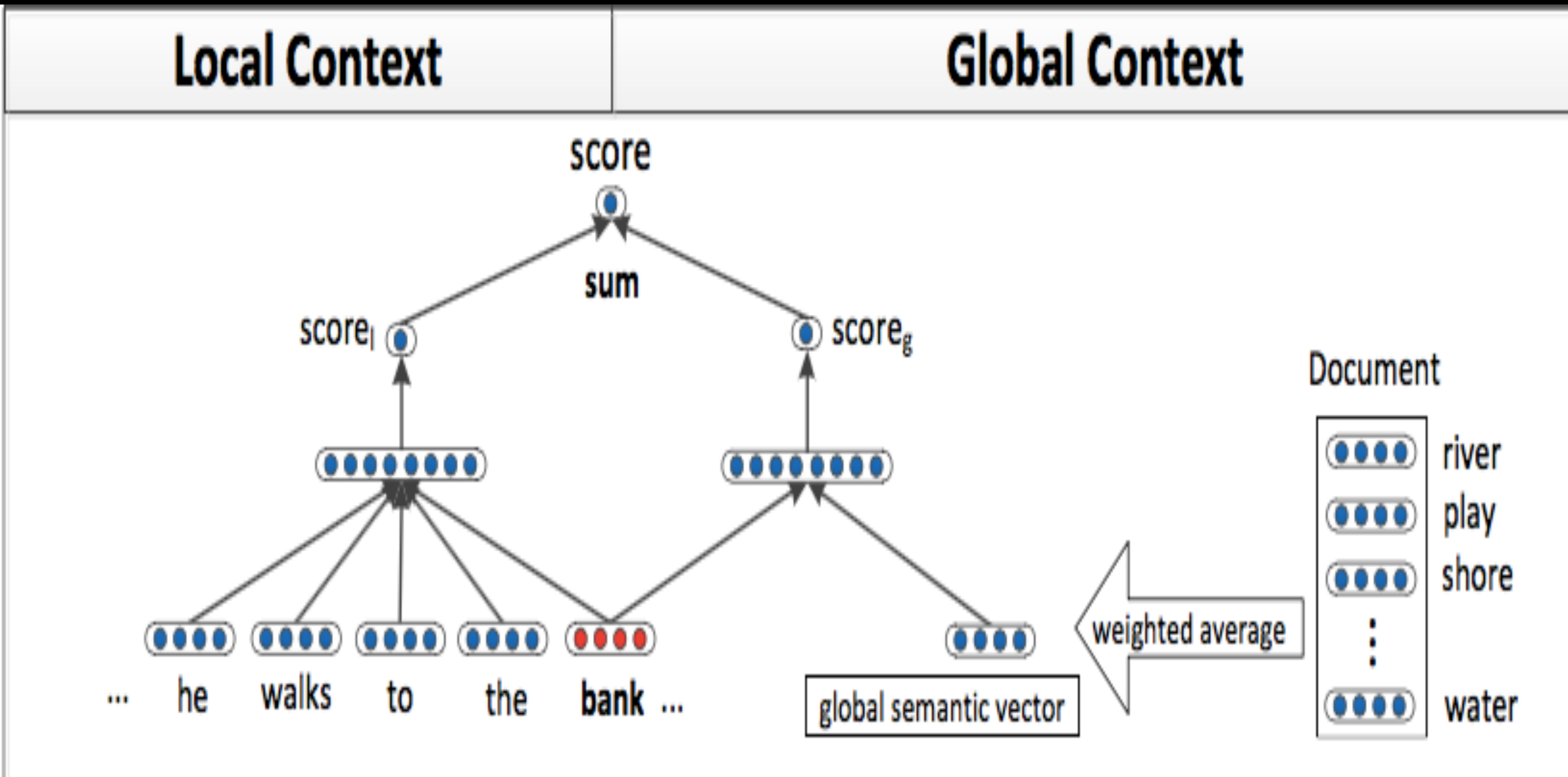
Neural Network Architecture

- two scoring components

1. local context

2. other global context

Local and global context



Score of local context

$$a_1 = f(W_1[x_1; x_2; ...; x_m] + b_1)$$

$$score_l = W_2 a_1 + b_2$$

x_i : embedding of word i in the sequence

Embedding matrix $L \in R^{n \times |V|}$

f: element-wise activation function

W_1, W_2: the first and second layer weights of the neural network

b_1, b_2: biases of each layer

Explanation

$$a_1 \in R^{h \times 1}$$

$$W_1 \in R^{h \times (mn)}$$

$$W_2 \in R^{1 \times h}$$

Score of the global context -1

$$c = \frac{\sum_{i=1}^{k} \omega(t_i) d_i}{\sum_{i=1}^{k} \omega(t_i)}$$

d_i: word embedding in document

$w(t_i)$: weight function that captures the importance of word t_i in the document.

Score of the global context-2

$$a_1^{(g)} = f(W_1^{(g)}[c;x_m] + b_1^{(g)})$$

$$score_g = W_2^{(g)} a_1^{(g)} + b_2^{(g)}$$

x_m: the last word in the word sequence
c: weighted average of all word vectors
 in the document

Explanation

$$a_1^{(g)} \in R^{h^{(g)} \times 1}$$

$$W_1^{(g)} \in R^{h^{(g)} \times (2n)}$$

$$W_2^{(g)} \in R^{1 \times h^{(g)}}$$

Final score

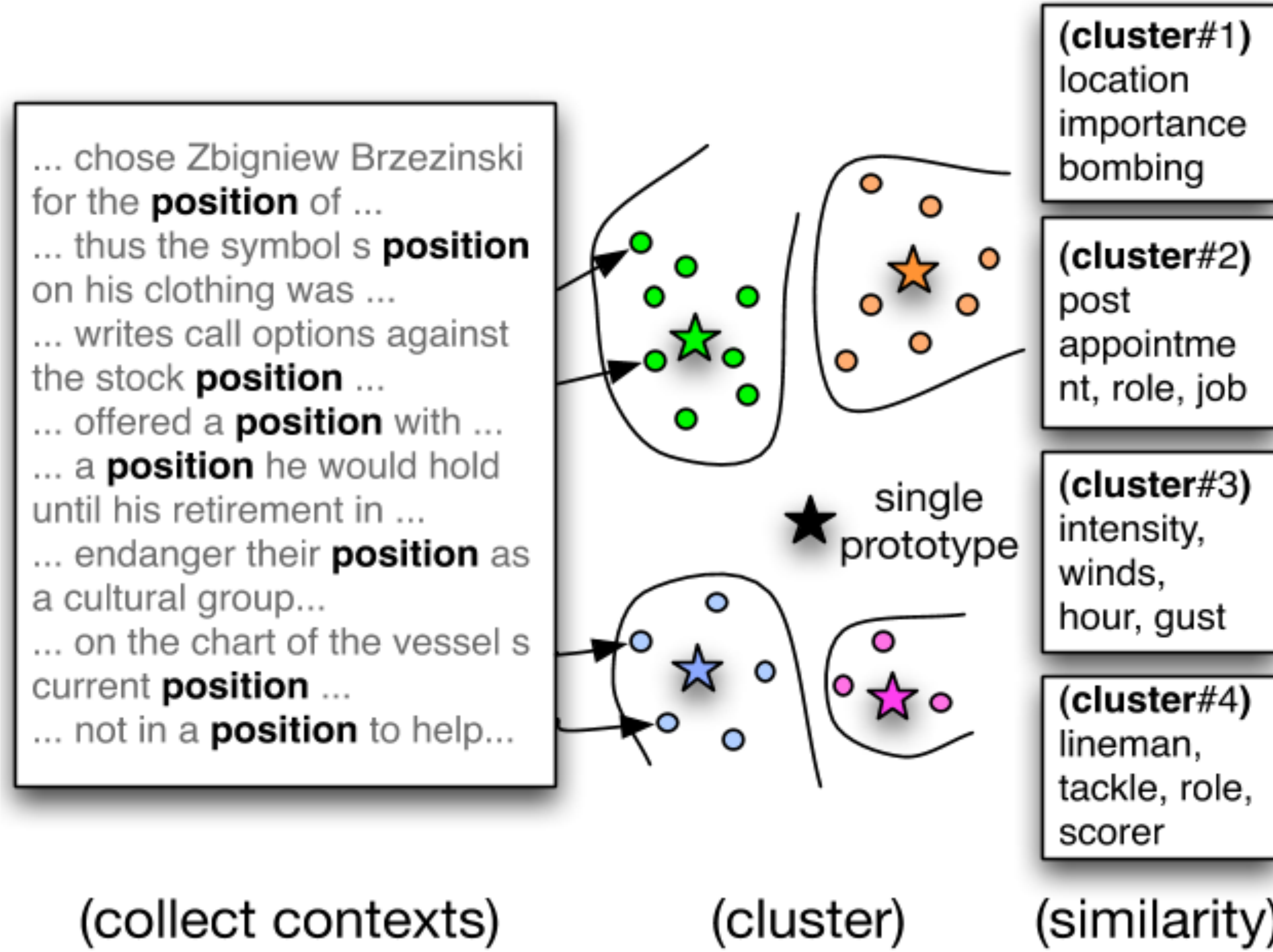
$$score = score_l + score_g$$

Now we has the score of a sequence-document pair —— g(s, d)

For corrupt examples, just randomly choosing a word from dictionary.

Multi-Prototype Neural Language Model

- use our learned single-prototype embeddings to represent each con- text window, which can then be used by clustering to perform word sense discrimination
- fixed-sized context windows of all occurrences of a word (5 words before and after the word occurrence)


... chose Zbigniew Brzezinski for the position of ...
... thus the symbol s position on his clothing was ...
... writes call options against the stock position ...
... offered a position with ...
... a position he would hold until his retirement in ...
... endanger their position as a cultural group...
... on the chart of the vessel s current position ...
... not in a position to help...
single prototype
(cluster#1) location importance bombing
(cluster#2) post appointme nt, role, job
(cluster#3) intensity, winds, hour, gust
(cluster#4) lineman, tackle, role, scorer
(collect contexts)
(cluster)
(similarity)

Clustering

- Each context is represented by a weighted average of the context words’ vectors (use idf-weighting as the weighting function)
- use spherical k-means to cluster these context representations and each word occurrence in the corpus is re-labeled to its associated cluster

Similarity between a pair of words

$$AvgSimC(\omega,\omega')=$$

$$\frac{1}{K^2}\sum_{i=1}^{k}\sum_{j=1}^{k}p(c,\omega,i)p(c',\omega',j)d(\mu_i(\omega),\mu_j(\omega'))$$

p(c,w,i): the likelihood of word w in the cluster i given context c

$\mu_i(\omega)$: the vector representing the i-th cluster centroid of w

Experiments

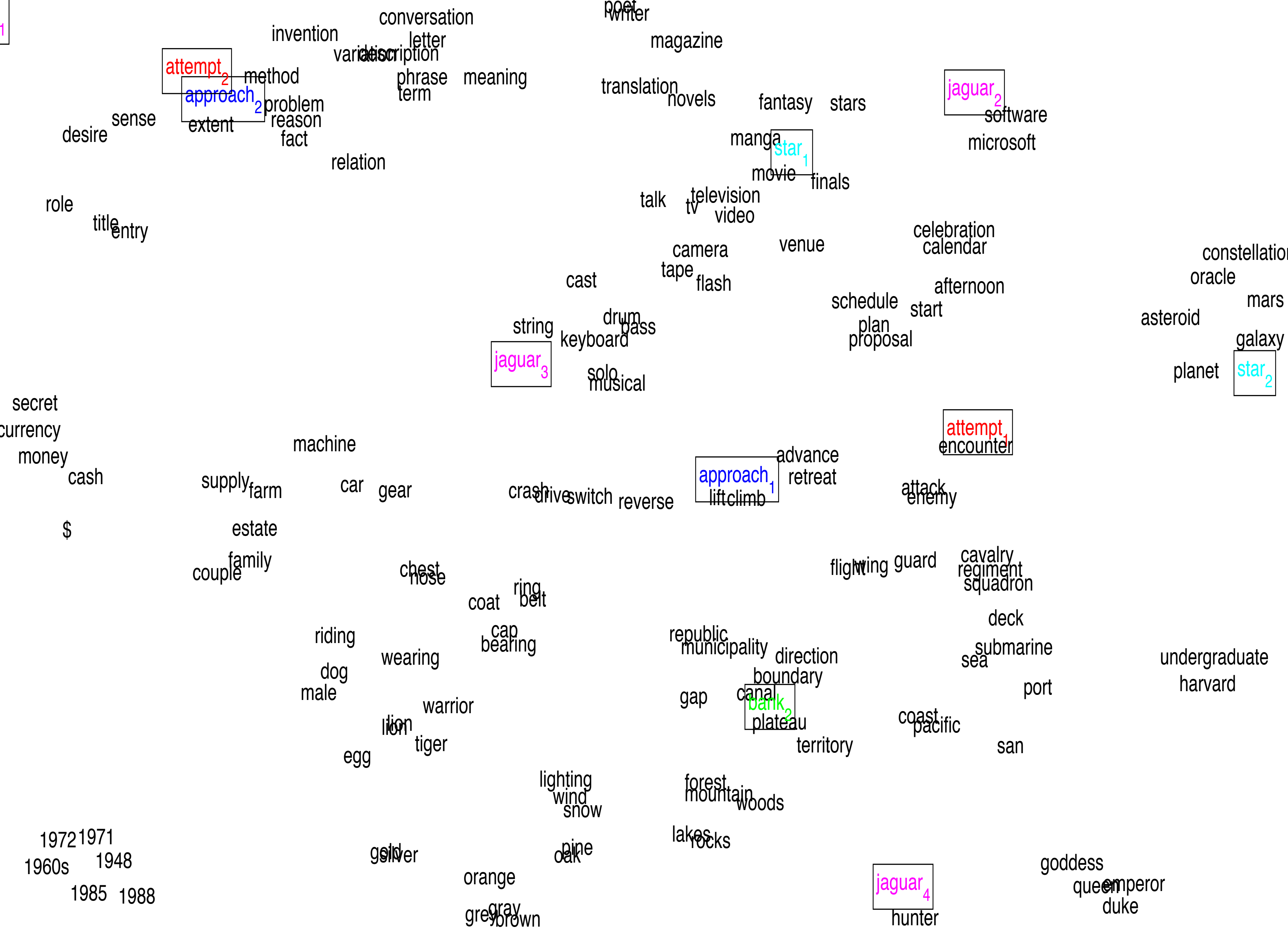
- 50- dimensional embeddings
- 10-word windows of text as the local context
- 100 hidden units
- no weight regularization for both neural networks
- fix the number of prototypes to be 10.

With global context

Center Word	C&W	Our Model
markets	firms, industries, stores	market, firms, businesses
American	Australian, Indian, Italian	U.S., Canadian, African
illegal	alleged, overseas, banned	harmful, prohibited, convicted

With multi-prototype

Center Word	Nearest Neighbors
bank_1	corporation, insurance, company
bank_2	shore, coast, direction
star_1	movie, film, radio
star_2	galaxy, planet, moon
cell_1	telephone, smart, phone
cell_2	pathology, molecular, physiology
left_1	close, leave, live
left_2	top, round, right


1
conversation
invention
letter
variety
description
attempt$_2$
method
approach$_2$
problem
reason
fact
extent
sense
desire
phrase
term
meaning
relation
poet
writer
magazine
translation
novels
fantasy
stars
jaguar$_2$
software
microsoft
manga
star$_1$
movie
finals
talk
television
tv
video
role
title
entry
celebration
calendar
camera
tape
flash
venue
cast
constellation
oracle
afternoon
schedule
start
plan
proposal
mars
asteroid
string
drum
bass
keyboard
jaguar$_3$
solo
musical
galaxy
planet
star$_2$
secret
currency
money
cash
machine
attempt$_1$
encounter
advance
approach$_1$
retreat
lift
climb
supply
farm
car
gear
crash
drive
switch
reverse
attack
enemy
$
estate
family
couple
chest
nose
ring
belt
coat
flight
wing
guard
cavalry
regiment
squadron
deck
riding
cap
bearing
republic
municipality
direction
submarine
sea
undergraduate
harvard
wearing
dog
male
boundary
gap
canal
bank$_2$
plateau
port
warrior
lion
lion
tiger
egg
coast
pacific
territory
san
lighting
wind
snow
forest
mountain
woods
1972
1971
1960s
1948
1985
1988
gold
silver
pine
oak
lakes
rocks
orange
grey
gray
brown
jaguar$_4$
hunter
goddess
queen
emperor
duke

Spearman's correlation on new dataset

Model	$\rho \times 100$
C&W-S	57.0
Our Model-S	58.6
Our Model-M AvgSim	62.8
Our Model-M AvgSimC	**65.7**

Q&A