Corpus and Linguistic Analysis

Weiwei Sun

Institute of Computer Science and Technology
Peking University

December 19, 2017

Corpus Linguistics

vs.

Corpus and Linguistic Analysis

Corpora are useful

A practical definition

> *A corpus provides texts in form of linguistically meaningful and retrievable units in a reusable way.*

(from Kübler & Zinsmeister, 2014, *Corpus Linguistics and Linguistically Annotated Corpora*)

Corpora serve

- collection of examples for linguists
- data resource for lexicographers
- instruction material for language teachers and learners
- natural language processing (NLP) applications
- linguistic analysis

Data

`http://bcc.blcu.edu.cn/`

Second language acquisition

- It pays to wait.
- It waits to pay.

What can we learn from `errors`?

Distributional Semantics

Natural language text = Sequences of words.

$\Downarrow$

How to represent words?

Naive representation

- The vast majority of rule-based and statistical NLP work regards words as atomic symbols:

  BBS, PKU, study

- Using vector space terms, this is a vector with one 1 and a lot of zeroes

  [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] in $\mathbb{R}^{|\text{vocabulary}|}$.

- Dimensionality is very large: 50K (PTB), 13M (Google 1T)

Natural language text = Sequences of words.

$$\Downarrow$$

How to represent words?

Naive representation

- The vast majority of rule-based and statistical NLP work regards words as atomic symbols:

  BBS, PKU, study

- Using vector space terms, this is a vector with one 1 and a lot of zeroes

  $[0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0]$ in $\mathbb{R}^{|\text{vocabulary}|}$.

- Dimensionality is very large: 50K (PTB), 13M (Google 1T)

- *fast* is similar to *rapid*
- *tall* is similar to *height*

Question answering

Q How tall is Mt. Everest?

Candidate A The official height of Mount Everest is 29029 feet

Guiding hypotheses

John Firth, (1957, *A synopsis of linguistic theory*)

> *You shall know a word by the company it keeps.*

> *the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously.*

Zellig Harris (1954, *Distributional structure*)

> *distributional statements can cover all of the material of a language without requiring support from other types of information.*

Idea

- To produce dense vector representations based on the context/use of words.
- Three main approaches: count-based, predictive, and task-based.

Count-based methods

- Define a basis vocabulary $\mathcal{C}$ of context words.
- Define a word window size w.
- Count the basis vocabulary words occurring w words to the left or right of each instance of a target word in the corpus.
- Form a vector representation of the target word based on these counts.

Corpus

`... and the cute kitten purred and then ...`

`... the cute furry cat purred and miaowed ...`

`... that the small kitten miaowed and she ...`

`... the loud furry dog ran and bit ...`

...

Example basis vocabulary:

{..., bit, cute, furry, loud, miaowed, purred, ran, small, ...}.

kitten context words: {cute, purred, small, miaowed, ...}.

cat context words: {furry, purred, ...}.

dog context words: {furry, ran, ...}.

Corpus

... `and the cute kitten purred and then` ...

... `the cute furry cat purred and miaowed` ...

... `that the small kitten miaowed and she` ...

... `the loud furry dog ran and bit` ...

...

Example basis vocabulary:

{..., bit, cute, furry, loud, miaowed, purred, ran, small, ...}.

> **kitten** context words: {cute, purred, small, miaowed, ...}.
>
> **cat** context words: {furry, purred, ...}.
>
> **dog** context words: {furry, ran, ...}.

An example

Corpus

```
... and the cute kitten purred and then ...
... the cute furry cat purred and miaowed ...
... that the small kitten miaowed and she ...
... the loud furry dog ran and bit ...
...
```

Example basis vocabulary:

$\{$..., bit, cute, furry, loud, miaowed, purred, ran, small, ...$\}$.

kitten context words: $\{$cute, purred, small, miaowed, ...$\}$.

cat context words: $\{$furry, purred, ...$\}$.

dog context words: $\{$furry, ran, ...$\}$.

Corpus

... and the cute kitten purred and then ...

... the cute furry cat purred and miaowed ...

... that the small kitten miaowed and she ...

... the loud furry dog ran and bit ...

...

Example basis vocabulary:

{..., bit, cute, furry, loud, miaowed, purred, ran, small, ...}.

kitten context words: {cute, purred, small, miaowed, ...}.

cat context words: {furry, purred, ...}.

dog context words: {furry, ran, ...}.

Corpus

`... and the cute kitten purred and then ...`

`... the cute furry cat purred and miaowed ...`

`... that the small kitten miaowed and she ...`

`... the loud furry dog ran and bit ...`

...

Example basis vocabulary:

{..., bit, cute, furry, loud, miaowed, purred, ran, small, ...}.

kitten context words: {cute, purred, small, miaowed, ...}.

cat context words: {furry, purred, ...}.

dog context words: {furry, ran, ...}.

An example

Corpus

```
... and the cute kitten purred and then ...
... the cute furry cat purred and miaowed ...
... that the small kitten miaowed and she ...
... the loud furry dog ran and bit ...
...
```

Example basis vocabulary:

{..., bit, cute, furry, loud, miaowed, purred, ran, small, ...}.

kitten$= [0, 1, 0, 0, 1, 1, 0, 1]^\top$

cat$= [0, 1, 1, 0, 1, 0, 0, 0]^\top$

dog$= [1, 0, 1, 1, 0, 0, 1, 0]^\top$

Problem with raw counts

Raw word frequency is not a great measure of association between words

> *the* and *of* are very frequent, but maybe not the most discriminative

Pointwise mutual information

Information-theoretic measurement: Do events x and y co-occur more than if they were independent?

$$PMI(X,Y) = \log \frac{P(x,y)}{P(x) \cdot P(y)}$$

	computer	data	pinch	result	sugar	
apricot	0	0	1	0	1	
pineapple	0	0	1	0	1	
digital	2	1	0	1	0	
information	1	6	0	4	0	

	computer	data	pinch	result	sugar	
apricot	0.00	0.00	0.05	0.00	0.05	
pineapple	0.00	0.00	0.05	0.00	0.05	
digital	0.11	0.05	0.00	0.05	0.00	
information	0.05	0.32	0.00	0.21	0.00	

	computer	data	pinch	result	sugar	p(word)
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58

	computer	data	pinch	result	sugar	p(word)
apricot	0.00	0.00	0.05	0.00	0.05	0.11
pineapple	0.00	0.00	0.05	0.00	0.05	0.11
digital	0.11	0.05	0.00	0.05	0.00	0.21
information	0.05	0.32	0.00	0.21	0.00	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

- Matrix: words $\times$ contexts
- f_{ij} is # of times w_i occurs in context c_j

	computer	data	pinch	result	sugar	p(word)
apricot			2.25		2.25	0.11
pineapple			2.25		2.25	0.11
digital	1.66	0.00		0.00		0.21
information	0.00	0.57		0.00		0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

- Matrix: words $\times$ contexts
- f_{ij} is # of times w_i occurs in context c_j

Problems

- PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Solution: Laplace (add-one) smoothing

	computer	data	pinch	result	sugar	
apricot	2	2	3	2	3	
pineapple	2	2	3	2	3	
digital	2	3	2	3	2	
information	3	8	2	6	2	

Problems

- PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Solution: Laplace (add-one) smoothing

	computer	data	pinch	result	sugar	
apricot	0.03	0.03	0.05	0.03	0.05	
pineapple	0.03	0.03	0.05	0.03	0.05	
digital	0.11	0.05	0.03	0.05	0.03	
information	0.05	0.14	0.03	0.10	0.03	

Problems

- PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Solution: Laplace (add-one) smoothing

	computer	data	pinch	result	sugar	p(word)
apricot	0.03	0.03	0.05	0.03	0.05	0.20
pineapple	0.03	0.03	0.05	0.03	0.05	0.20
digital	0.11	0.05	0.03	0.05	0.03	0.24
information	0.05	0.14	0.03	0.10	0.03	0.36

Weighting PMI

Problems

- PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Solution: Laplace (add-one) smoothing

	computer	data	pinch	result	sugar	p(word)
apricot	0.03	0.03	0.05	0.03	0.05	0.20
pineapple	0.03	0.03	0.05	0.03	0.05	0.20
digital	0.11	0.05	0.03	0.05	0.03	0.24
information	0.05	0.14	0.03	0.10	0.03	0.36
p(context)	0.19	0.25	0.17	0.22	0.17	

Problems

- PMI is biased toward infrequent events
- Very rare words have very high PMI values
- Solution: Laplace (add-one) smoothing

	computer	data	pinch	result	sugar	p(word)
apricot			0.56		0.56	0.20
pineapple			0.56		0.56	0.20
digital	0.62	0.00		0.00		0.24
information	0.00	0.58		0.37		0.36
p(context)	0.19	0.25	0.17	0.22	0.17	

Using syntax to define a word's context

Zelig Harris (1968)

> *The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities.*

- Two words are similar if they have similar syntactic contexts
- duty and responsibility have similar syntactic distribution:
  - **Modified by adjectives**: additional, administrative, assumed, collective, congressional, constitutional, ...
  - **Objects of verbs**: assert, assign, assume, attend to, avoid, become, breach, ...

Context based on dependency parsing (1)

`I have a brown dog`

(have subj I), (I subj-of have), (dog obj-of have), (dog adj-mod brown), (brown adj-mod-of dog), (dog det a), (a det-of dog)

The description of *cell*

count(cell, subj-of, absorb)=1
count(cell, subj-of, adapt)=1
count(cell, subj-of, behave)=1
...
count(cell, pobj-of, in)=159
count(cell, pobj-of, inside)=16
count(cell, pobj-of, into)=30
...

Given two target words, we'll need a way to measure their similarity.

- Take angle between vectors as measure of similarity.
  - (correctly) ignores length of vectors = frequency of words
  - similar angle = similar proportion of context words
- Cosine of angle is easy to compute.
  - $\cos = 1$ means angle is 0°, i.e. very similar
  - $\cos = 0$ means angle is 90°, i.e. very dissimilar

$$\begin{aligned}\cos(u, v) &= \frac{u^\top v}{||u|| \cdot ||v||} \\ &= \frac{\sum_{i=1}^{n} u_i \cdot v_i}{\sqrt{\sum_{i=1}^{n} u_i \cdot u_i} \cdot \sqrt{\sum_{i=1}^{n} v_i \cdot v_i}}\end{aligned}$$

Many other methods to compute similarity

Context based on dependency parsing (2)

hope (N):
optimism 0.141, chance 0.137, expectation 0.136, prospect 0.126, dream 0.119, desire 0.118, fear 0.116, effort 0.111, confidence 0.109, promise 0.108

hope (V):
would like 0.158, wish 0.140, plan 0.139, say 0.137, believe 0.135, think 0.133, agree 0.130, wonder 0.130, try 0.127, decide 0.125

brief (N):
legal brief 0.139, affidavit 0.103, filing 0.098, petition 0.086, document 0.083, argument 0.083, letter 0.079, rebuttal 0.078, memo 0.077, article 0.076

brief (A):
lengthy 0.256, hour-long 0.191, short 0.173, extended 0.163, frequent 0.162, recent 0.158, short-lived 0.155, prolonged 0.149, week-long 0.149, occasional 0.146

Reference

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words.

Similarity = synonymy?

- Antonyms are basically as distributionally similar as synonyms:
- Distributional similarity is not referential similarity.
- Distinguishing synonyms from antonyms is notoriously hard problem.

brief (A):
lengthy 0.256, hour-long 0.191, short 0.173, extended 0.163, frequent 0.162, recent 0.158, short-lived 0.155, prolonged 0.149, week-long 0.149, occasional 0.146