Natural Language Processing

Assignment 1 – Report

Tommy Yu

Apr. 10, 2018

1.1 Distributional Counting

C - Spearman correlation (w = 3)	
MEN	0.2281
SimLex-999	0.0561

1.2 Computing PMIs

C_{PMI} - Spearman correlation (w = 3)	
MEN	0.5331
SimLex-999	0.2238

1.3 Experimentation

Spearman correlation	w = 1		w = 3		w = 6	
	MEN	SimLex-999	MEN	SimLex-999	MEN	SimLex-999
C	0.2062	0.0751	0.2281	0.0561	0.2302	0.0332
C_{PMI}	0.4640	0.2679	0.5331	0.2238	0.5267	0.1751

Trends: For both C and C_{PMI}, as window size increases, the correlation with MEN dataset tends to increase, but seemingly at a decreasing rate. It largely remains at the same level between w=3 and w=6; however, the correlation with SimLex-999 dataset steadily decreases.

By examining the datasets manually, we can see that SimLex-999 dataset mainly assigns higher scores to synonyms, and lower scores to antonyms, as well as pairs of words not close in meaning, yet still sharing the part-of-speech tags. As a result, it is more likely to identify identical words amongst closer neighbors, e.g., for synonyms which can usually replace one another: “smart person” and “intelligent person”. When the window size increases, more “noise” from other parts of the sentence is introduced, and thus the correlation deviates from the human-annotated scores.

In contrast, MEN dataset assigns higher scores to a pair words in the same context, like “camera” and “photography” – yet the two words have very different meaning, and would appear in a sentence in very different ways. Therefore, increasing the window size to a certain extent can better capture the context of the center word, therefore increasing the correlation, before too much “noise” is introduced.

1.4.1 Printing Nearest Neighbors

10 nearest neighbors for "monster"

w = 1	w = 6
dragon	evil
tyrant	giant
creatures	creature
monsters	monsters
jar	godzilla
hornet	dragon
invaders	dog
rhinoceros	ghost
robot	girl
gangster	horror

1.4.2 POS Tag Similarity

Applying the pos_tag function from the nltk library to the 25,000 most common words in Wikipedia yields the following counts:

Nouns	18991
Verbs	2795
Adjectives	1606
Prepositions	58
Others	1550

50 words each are randomly selected from nouns, verbs, adjectives, and prepositions. The tagging of some words may not be accurate, but it should not affect the examination of the general trend. The average percentage of different tags for each of the four categories is summarized as follows.

10 Nearest Neighbors	Noun		Verb		Adjective		Preposition		Others	
	w=1	w=6	w=1	w=6	w=1	w=6	w=1	w=6	w=1	w=6
Noun	89%	86%	5%	7%	4%	5%	0%	0%	2%	1%
Verb	35%	69%	56%	19%	6%	7%	1%	1%	3%	4%
Adjective	49%	68%	14%	8%	33%	18%	0%	0%	4%	5%
Preposition	12%	32%	12%	9%	4%	7%	37%	19%	35%	33%

In general, there is a systematic decrease in part-of-speech similarity in nearest neighbors when the window size is increased from 1 to 6, as shown in the green cells. The decrease in very apparent for verbs, adjectives, and prepositions, yet not so much for nouns, as even with w=6, nouns still account for the vast majority of the nearest neighbors.

Again, nouns are comparable to nouns most of the time. Take "consequence" as an example,

w=1: `['mistake', 'tendency', 'interruption', 'devastation', 'outcome', 'auspices', 'implications', 'continuation', 'consequences', 'conclusion']`
w=6: `['risk', 'behaviour', 'affect', 'measures', 'cause', 'depends', regarding', 'outcomes', 'circumstances', 'patients']`

Although there is no common neighbor across the two window sizes, most neighbors remain nouns after increasing w. Another thing to note is that when window size increases, some obvious synonyms/replacement words may disappear from the list due to the "noise" brought in when w is

too large. For example, "consequences" is no longer in the top 10 when w=6. And this is a very common phenomenon across the words examined.

For verbs, adjectives, and prepositions, when the window size increases from 1 to 6, it appears that nouns are making up for the decrease in nearest neighbors of the same POS tag (see the red cells). One example from each is provided; nouns for w=6 are in red.

"questionable" (adj.)

w=1: ['dubious', 'doubtful', 'problematic', 'biased', 'controversial', 'debatable', 'tendentious', 'unverified', 'unclear', 'encyclopedic']

---mostly adjectives

w=6: ['facts', 'citations', 'verify', 'verifiable', 'dubious', 'unsourced', 'questioned', 'opinion', 'npov', 'sourced']

"modifying" (v.)

w=1: ['referring', 'masculine', '1695', 'deleting', 'vandalizing', 'redirecting', 'entirety', 'restoring', 'recreating', 'synthesized']

---mostly verbs

w=6: ['firmware', 'indefinite', 'emissions', 'wikilink', 'specifies', 'update', 'ethernet', 'hardware', 'reverting', 'vandals']

"during" (prep.)

w=1: ['and', 'were', 'after', 'from', 'when', 'at', 'between', 'until', 'had', 'before']

---mostly prepositions

w=6: ['war', 'after', 'early', 'years', 'until', 'year', 'following', 'before', 'world', 'later']

It turns out that when the window size increases, nouns that are frequently used in close vicinity of a v./adj./prep. center word will more likely share similar neighbors with the center word itself.

In general, when the window size increases, many nearest neighbors get replaced. Of the 200 words randomly chose, very few words share over 5 out of 10 nearest neighbors between the two window sizes. "until" and "southeastern" witnesses the most common neighbors, with 9 and 8 out of 10. Those in black are common neighbors.

"until" (prep.)

w=1: ['december', 'october', 'november', 'january', 'april', 'june', 'february', 'july', 'september', 'august']
w=6: ['january', 'march', 'june', 'july', 'september', 'august', 'october', 'december', 'april', 'november']

Probably due to Wikipedia's tendency to use "until" over "till" when formally describing a time period.

"southeastern" (adj.)

w=1: ['northeastern', 'northwestern', 'southwestern', 'eastern', 'southern', 'northern', 'western', 'southwest', 'south-central', 'southeast']

w=6: ['northeastern', 'southwestern', 'southern', 'northeast', 'northwestern', 'southwest', 'eastern', 'western', 'northern', 'northwest']
This is very straightforward – all directions.

1.4.3 Word with Multiple Senses

In this section, 30 common words with multiple senses are examined:

```
"foot", "bill", "bit", "bat", "star", "seal", "can", "club", "bank",
"bear", "pool", "pound", "head", "bore", "current", "custom", "doctor",
"channel", "novel", "patient", "plane", "strike", "like", "charge",
"minor", "suit", "trace", "chair", "company", "date"
```

As expected, there is not much systematic trend in general, as there can be great variance in terms of how frequent different senses are used across these words in Wikipedia. Of the 30 words examined, here are a few examples that representing certain patterns the best.

(1) One sense dominates the rest despite the change in window size.

bill – "Bill" as a name dominates other senses such as "a form of currency", or "draft of law".

```
w=1: ['john', 'william', 'david', 'jim', 'james', 'george', 'tom', 'mike',
 'robert', 'bob']
w=6: ['david', 'james', 'john', 'bob', 'william', 'michael', 'robert', 'george', 'smith', 'paul']
```

(2) More senses are brought up as window size increases.

bank – the sense as "riverbank" shows up when w=6.

```
w=1: ['banks', 'insurance', 'company', 'corporation', 'banking', 'government', 'railway', 'library', 'business', 'companies']
w=6: ['banks', 'river', 'corporation', 'west', 'capital', 'company', 'railway', 'located', 'central', 'east']
```

(3) Condensing towards the word's most popular usage as window size increases.

channel – not so intuitive neighbors like "minutes" or "seconds" when w=1; yet almost entirely referring to "media channels" when w=6.

```
w=1: ['channels', 'network', 'television', 'tv', 'percent', 'satellite',
'route', 'minutes', 'tackles', 'seconds']
w=6: ['television', 'tv', 'network', 'broadcast', 'radio', 'broadcasting',
 'news', 'channels', 'aired', 'cable']
```