NLP Exercise 3

Eyal Orbach - id 015369317
Daniel Juravski - id 206082323

Similarity Computation:

Some Statistics:

We evaluated the similarity computations with 2 main filters:

1. We filtered words that occurred less than **100** times in the corpus.
2. We filtered features that occurred with the target word less than **5** times in the corpus.
   We removed this limitation for the 'Dependency' strategy since the initial results were less accurate then expected
   (producing about 6 irrelevant words in the top 20). Removing this filter successfully improved the results of this strategy.
   Prior to that we experimented with smoothing the PMI which did not yield improvement here.

After using those filters, the matrix dimensions were as follow:

	Type1: All-sentence	*Type2: Window*	*Type3: Dependency*
Number of words	7,947	7,947	7,776
Number of featues	51,555	19,829	1,278,171
Avg. features for word	314	69	854
Max. features for word	45,098	15,984	136,932
Min. features for word	3	2	1

conclusions:

Full tables are in the included appendix files, in table form at ASCI Format (.txt)

As described above we have removed the threshold for 'word context co-occurrence' for the dependency co-occurrence strategy, which improved the results significantly.
Initially the dependency top contexts were more general (car->damage) but after the removal they became incredibly unique and distinct (cat->firmly-sprung)

2. similar words:

- Type1: All-sentence: In this strategy, we'll get words that are strongly related to the topic of the target word, but their syntactic roles are sometimes very different, meaning they cannot replace the target word in a sentence . Example: *hospital* -> [health, patient, surgery].
- Type2: Window: In this strategy, there appears to be some mix between the topic and the semantic class, while some word share the semantic class but more loosely relate to the topic (*facility*) others will share the topic but be of different syntactic roles. Example: *hospital -> [illness, facility, center].*
- Type3: Dependency: In this strategy, the words will usually share a semantic class with the target word and be syntacticly similar, meaning they can replace the target word in a sentence, but they might relate more loosely to the target's topic, and have very different meaning example: *hospital -> [museum, university, prison*].

3. top context attributes:

Full tables are in the included appendix files, in table form at ASCI Format (.txt)

In these lists we can find many words that describe the target word, or its usage

gun->large-caliber, guitar-> acoustic , bomb -> explode

and also some specific models/companies that describe the target word

hotel->hilton, car->jaguar

Adjectives and verbs that relate to the target word are popular contexts but rarely came up as a similar word for a noun target word

4. MAP results

AP calculations can be at the end of this PDF. We used N to be a the sum of the union of all relevant words found by the three co-occurrence types.
MAP results:

sentence = (0.53 + 0.645) /2 = 0.588

window = (0.575 + 0.645) / 2 = 0.61

dependency = (0.567 + 0.645) / 2 = 0.606

The window based co-occurrence is the winner here and we attribute it to its mixture of topical and syntactic relations.

5.Implementation Description:

- Estimation of PMI values:
  - We used the following formula $PMI(a,b) = \log \frac{P(a,b)}{P(a) \cdot P(b)}$
  - For that we had calculated these probabilities:
    - P(a,b) = P(word, attribute), is # of occurrence of word a with the attribute b, over all the possible combinations of any word and attribute in the corpus.
    - P(a) = P(word), is # of occurrence of word a, over all the other words occurrences in the corpus.
    - P(b) = P(attribute), is # of occurrence of attribute b, over all the other attributes occurrences in the corpus.
  - We did the sanity checks and made sure that each one of the probabilities is equal to **1**
  - When we got negative PMI value, we set that value to 0.
  - We have tried both including negative values, and smoothing (see similarity_computation.py line 91) but eventually set the smoothing factor back to 1 since we saw no significant improvement.
  - The PMI code is in 'similarity_computation.py' file, 'PMI' function (lines 18-28).

- Efficient algorithm:
  - The algorithm is based on computing the PMI values over common attributes only.
  - We had an extra dict with mapping of attributes to words. Thus, when we got word u with its attributes, we computed only u's attributes, only over the words that had those attributes.
  - Complexity reduced by "sparseness" factor.
  - The efficient algorithm code is in 'similarity_computation.py' file, 'findSimilar' function (lines 115-133).

6. WordToVec

6.2 **word similarity**

Full tables are in the included appendix files, in table form at ASCI Format (.txt)

BOW5: In this strategy we saw often ‘specific implementation’ of the target word (table->billiard, bus -> inter-city, gun->sub-machine) either nouns or adjectives that describe the target word but have little chance of describing an object different from the target word (guitar-> fretless), and therefore have high likelihood of being adjacent to the target word and perhaps even replacing it sometimes (“…buy inter-city tickets…”)

Dependency: In this strategy we again noticed words that are more loosely related to the topic of the target word but share a semantic class and could syntactically replace the target word in many contexts but that would cause a significant change in the meaning (bus->ferryboat, hospital->guesthouse, horse->elephant)

6.3 Top context attributes

Full tables are in the included appendix files, in table form at ASCI Format (.txt)

BOW5: In this list we saw several of the nouns/adjectives described above that appeared in the closest words list (gun->sub-machine) but also some nouns, adjectives and verbs that are common for the target word but are not exclusive to it
(car-dealership, horse->riding, bomb->atomic)

Dependency: In this list we also found several non-exclusive nouns, verbs and adjectives that describe the target word. *compmod* and *adpmod* were the dominant relations to the target word

6.4 MAP results

AP calculations can be at the end of this PDF.
We used N to be a the sum of the union of all relevant words found by the the 2 embedding strategies of word2vec.
MAP results:

bow5 = (0.654 + 0.741) /2 = 0.698

dependency = (0.664 + 0.741) / 2 = 0.703

6. word2vec conclusions

Initially the results on dependency word2vec reached higher results than our own dependency co-occurrence strategy, which led us to improve ours by dropping the filter as described above. After that the percentage of related words was similar between our implementation and word2vec on the dependency and 5-ngram word window strategies, with the full sentence co-occurrence getting slightly lower results.

The MAP values are higher for word2vec since the union of the relevant words yield more results on our similarity finders, resulting in the N denominator being larger (this is true also if we only take into account 2 co-occurrence types and not all 3)

The fact that the embedding vectors produce fixed length vectors enables the calculation of finding a similar word to be much faster.

Semantic class and topical relation

car

#	Closest Words - car					
	Type1: all-sentence		Type2: Window		Type3: Dependency	
1	vehicle	Semantic: yes Topical: yes	vehicle	Semantic: yes Topical: yes	vehicle	Semantic: yes Topical: yes
2	driver	Semantic: no Topical: yes	driver	Semantic: no Topical: yes	truck	Semantic: yes Topical: yes
3	race	Semantic: no Topical: yes	train	Semantic: yes Topical: yes	automobile	Semantic: yes Topical: yes
4	drive	Semantic: no Topical: yes	racing	Semantic: no Topical: yes	motorcycle	Semantic: yes Topical: yes
5	racing	Semantic: no Topical: yes	bus	Semantic: yes Topical: yes	bus	Semantic: yes Topical: yes
6	motor	Semantic: no Topical: yes	passenger	Semantic: no Topical: yes	wagon	Semantic: yes Topical: yes
7	engine	Semantic: no Topical: yes	race	Semantic: no Topical: yes	boat	Semantic: yes Topical: yes
8	truck	Semantic: yes Topical: yes	auto	Semantic: no Topical: yes	engine	Semantic: no Topical: yes
9	wheel	Semantic: no Topical: yes	automobile	Semantic: yes Topical: yes	carriage	Semantic: yes Topical: yes
10	automobile	Semantic: yes Topical: yes	motorcycle	Semantic: yes Topical: yes	aircraft	Semantic: yes Topical: yes
11	model	Semantic: no Topical: yes	motor	Semantic: no Topical: yes	ship	Semantic: yes Topical: yes
12	passenger	Semantic: no Topical: yes	drive	Semantic: no Topical: yes	locomotive	Semantic: yes Topical: yes
13	train	Semantic: yes Topical: yes	formula	Semantic: no Topical: yes	bicycle	Semantic: yes Topical: yes
14	front	Semantic: no Topical: no	truck	Semantic: yes Topical: yes	train	Semantic: yes Topical: yes
15	motorcycle	Semantic: yes Topical: yes	nascar	Semantic: no Topical: yes	equipment	Semantic: no Topical: no
16	rear	Semantic: no Topical: no	traffic	Semantic: no Topical: yes	driver	Semantic: no Topical: yes
17	auto	Semantic: no Topical: no	aircraft	Semantic: yes Topical: yes	plane	Semantic: yes Topical: yes
18	ford	Semantic: no Topical: yes	ship	Semantic: yes Topical: yes	motor	Semantic: no Topical: yes
19	bmw	Semantic: no Topical: yes	gt	Semantic: no Topical: yes	bike	Semantic: yes Topical: yes
20	dodge	Semantic: no Topical: yes	run	Semantic: no Topical: no	tram	Semantic: yes Topical: yes

piano

#	*Closest Words - piano*					
	Type1: all-sentence		*Type2: Window*		*Type3: Dependency*	
1	violin	Semantic: yes Topical: yes	violin	Semantic: yes Topical: yes	violin	Semantic: yes Topical: yes
2	flute	Semantic: yes Topical: yes	cello	Semantic: yes Topical: yes	flute	Semantic: yes Topical: yes
3	cello	Semantic: yes Topical: yes	op	Semantic: no Topical: yes	cello	Semantic: yes Topical: yes
4	concerto	Semantic: no Topical: yes	solo	Semantic: no Topical: yes	viola	Semantic: yes Topical: yes
5	solo	Semantic: no Topical: yes	guitar	Semantic: yes Topical: yes	guitar	Semantic: yes Topical: yes
6	viola	Semantic: yes Topical: yes	bass	Semantic: yes Topical: yes	trumpet	Semantic: yes Topical: yes
7	string	Semantic: no Topical: yes	concerto	Semantic: no Topical: yes	saxophone	Semantic: yes Topical: yes
8	sonata	Semantic: no Topical: yes	viola	Semantic: yes Topical: yes	keyboard	Semantic: yes Topical: yes
9	op	Semantic: no Topical: yes	flute	Semantic: yes Topical: yes	bass	Semantic: yes Topical: yes
10	instrument	Semantic: yes Topical: yes	string	Semantic: no Topical: yes	percussion	Semantic: yes Topical: yes
11	percussion	Semantic: yes Topical: yes	acoustic	Semantic: no Topical: yes	drum	Semantic: yes Topical: yes
12	trumpet	Semantic: yes Topical: yes	orchestra	Semantic: no Topical: yes	organ	Semantic: yes Topical: yes
13	quartet	Semantic: no Topical: yes	instrument	Semantic: yes Topical: yes	instrument	Semantic: yes Topical: yes
14	saxophone	Semantic: yes Topical: yes	sonata	Semantic: no Topical: yes	choir	Semantic: no Topical: yes
15	bass	Semantic: yes Topical: yes	perform	Semantic: no Topical: yes	horn	Semantic: yes Topical: yes
16	horn	Semantic: yes Topical: yes	quartet	Semantic: no Topical: yes	vocal	Semantic: no Topical: yes
17	keyboard	Semantic: yes Topical: yes	saxophone	Semantic: yes Topical: yes	orchestra	Semantic: no Topical: yes
18	composition	Semantic: no Topical: yes	choir	Semantic: no Topical: yes	solo	Semantic: no Topical: yes
19	ensemble	Semantic: no Topical: yes	musical	Semantic: no Topical: yes	music	Semantic: no Topical: yes
20	tenor	Semantic: no Topical: yes	music	Semantic: no Topical: yes	jazz	Semantic: no Topical: yes

AP calculation

AP car

#	Closest Words											
	Type1: all-sentence	rel	prec	Σ(prec * rel)	***Type2: Window***	rel	prec	Σ(prec * rel)	***Type3: Dependency***	rel	prec	Σ(prec * rel)
1	vehicle	1	1	1	vehicle	1	1	1	vehicle	1	1	1
2	driver	1	1	2	driver	1	1	2	truck	1	1	2
3	race	1	1	3	train	1	1	3	automobile	1	1	3
4	drive	1	1	4	racing	1	1	4	motorcycle	1	1	4
5	racing	1	1	5	bus	1	1	5	bus	1	1	5
6	motor	1	1	6	passenger	1	1	6	wagon	1	1	6
7	engine	1	1	7	race	1	1	7	boat	1	1	7
8	truck	1	1	8	auto	1	1	8	engine	1	1	8
9	wheel	1	1	9	automobile	1	1	9	carriage	1	1	9
10	automobile	1	1	10	motorcycle	1	1	10	aircraft	1	1	10
11	model	1	1	11	motor	1	1	11	ship	1	1	11
12	passenger	1	1	12	drive	1	1	12	locomotive	1	1	12
13	train	1	1	13	formula	1	1	13	bicycle	1	1	13
14	front	0	13/14	13	truck	1	1	14	train	1	1	14
15	motorcycle	1	14/15	13.93	nascar	1	1	15	equipment	0	14/15	14
16	rear	0	14/16	13.93	traffic	1	1	16	driver	1	15/16	14.93
17	auto	1	15/17	14.82	aircraft	1	1	17	plane	1	16/17	15.88
18	ford	1	16/18	15.7	ship	1	1	18	motor	1	17/18	16.82
19	bmw	1	17/19	16.6	gt	1	1	19	bike	1	18/19	17.77
20	dodge	1	18/20	17.5	run	0	19/20	19	tram	1	19/20	18.72
N= 33 (union of all relevant words)		AP = 0.53			AP = 0.575				AP = 0.567			

AP Piano

#	Closest Words											
	Type1: all-sentence	rel	prec	Σ	Type2: Window	rel	prec	Σ	Type3: Dependency	rel	prec	Σ
1	violin	1	1	1	violin	1	1	1	violin	1	1	1
2	flute	1	1	2	cello	1	1	2	flute	1	1	2
3	cello	1	1	3	op	1	1	3	cello	1	1	3
4	concerto	1	1	4	solo	1	1	4	viola	1	1	4
5	solo	1	1	5	guitar	1	1	5	guitar	1	1	5
6	viola	1	1	6	bass	1	1	6	trumpet	1	1	6
7	string	1	1	7	concerto	1	1	7	saxophone	1	1	7
8	sonata	1	1	8	viola	1	1	8	keyboard	1	1	8
9	op	1	1	9	flute	1	1	9	bass	1	1	9
10	instrument	1	1	10	string	1	1	10	percussion	1	1	10
11	percussion	1	1	11	acoustic	1	1	11	drum	1	1	11
12	trumpet	1	1	12	orchestra	1	1	12	organ	1	1	12
13	quartet	1	1	13	instrument	1	1	13	instrument	1	1	13
14	saxophone	1	1	14	sonata	1	1	14	choir	1	1	14
15	bass	1	1	15	perform	1	1	15	horn	1	1	15
16	horn	1	1	16	quartet	1	1	16	vocal	1	1	16
17	keyboard	1	1	17	saxophone	1	1	17	orchestra	1	1	17
18	composition	1	1	18	choir	1	1	18	solo	1	1	18
19	ensemble	1	1	19	musical	1	1	19	music	1	1	19
20	tenor	1	1	20	music	0	1	20	jazz	1	1	20
N= 31 (union of all relevant words)		AP = 0.645			AP = 0.645				AP = 0.645			

Word2Vec semantic class and topical relation

car

#	Closest Words - car			
	Word2Vec: bow5		Word2Vec: Dependency	
1	cars	Semantic: no Topical: yes	truck	Semantic: yes Topical: yes
2	truck	Semantic: yes Topical: yes	suv	Semantic: yes Topical: yes
3	automobile	Semantic: yes Topical: yes	vehicle	Semantic: yes Topical: yes
4	vehicle	Semantic: yes Topical: yes	minivan	Semantic: yes Topical: yes
5	motorbike	Semantic: yes Topical: yes	cars	Semantic: no Topical: yes
6	motorcycle	Semantic: yes Topical: yes	speedboat	Semantic: yes Topical: yes
7	driver	Semantic: no Topical: yes	racecar	Semantic: yes Topical: yes
8	minivan	Semantic: yes Topical: yes	automobile	Semantic: yes Topical: yes
9	suv	Semantic: yes Topical: yes	motorcar	Semantic: yes Topical: yes
10	lorry	Semantic: no Topical: no	jeep	Semantic: yes Topical: yes
11	motorcar	Semantic: yes Topical: yes	limousine	Semantic: yes Topical: yes
12	mid-engined	Semantic: no Topical: yes	minibus	Semantic: yes Topical: yes
13	limousine	Semantic: yes Topical: yes	lorry	Semantic: no Topical: no
14	front-engined	Semantic: no Topical: yes	limo	Semantic: yes Topical: yes
15	moped	Semantic: yes Topical: yes	motorcycle	Semantic: yes Topical: yes
16	motorhome	Semantic: yes Topical: yes	bike	Semantic: yes Topical: yes
17	mercedes-benz	Semantic: no Topical: yes	motorhome	Semantic: yes Topical: yes
18	bike	Semantic: yes Topical: yes	taxicab	Semantic: yes Topical: yes
19	rear-engined	Semantic: no Topical: yes	roadster	Semantic: no Topical: yes
20	three-wheeled	Semantic: no Topical: yes	wagon	Semantic: yes Topical: yes

piano

#	Closest Words - car			
	Word2Vec: bow5		Word2Vec: Dependency	
1	violin	Semantic: yes Topical: yes	violin	Semantic: yes Topical: yes
2	cello	Semantic: yes Topical: yes	cello	Semantic: yes Topical: yes
3	harpsichord	Semantic: yes Topical: yes	harpsichord	Semantic: yes Topical: yes
4	clarinet	Semantic: yes Topical: yes	saxophone	Semantic: yes Topical: yes
5	viola	Semantic: yes Topical: yes	clarinet	Semantic: yes Topical: yes
6	flute	Semantic: yes Topical: yes	guitar	Semantic: yes Topical: yes
7	bassoon	Semantic: yes Topical: yes	trombone	Semantic: yes Topical: yes
8	violoncello	Semantic: yes Topical: yes	mandolin	Semantic: yes Topical: yes
9	oboe	Semantic: yes Topical: yes	vibraphone	Semantic: yes Topical: yes
10	concerto	Semantic: no Topical: yes	marimba	Semantic: yes Topical: yes
11	saxophone	Semantic: yes Topical: yes	accordion	Semantic: yes Topical: yes
12	accordion	Semantic: yes Topical: yes	pianoforte	Semantic: yes Topical: yes
13	harp	Semantic: yes Topical: yes	bassoon	Semantic: yes Topical: yes
14	trombone	Semantic: yes Topical: yes	fortepiano	Semantic: yes Topical: yes
15	sonatas	Semantic: no Topical: yes	violoncello	Semantic: yes Topical: yes
16	trumpet	Semantic: yes Topical: yes	trumpet	Semantic: yes Topical: yes
17	mandolin	Semantic: yes Topical: yes	harmonica	Semantic: yes Topical: yes
18	pianoforte	Semantic: yes Topical: yes	clavinet	Semantic: yes Topical: yes
19	vibraphone	Semantic: yes Topical: yes	clavichord	Semantic: yes Topical: yes
20	concertos	Semantic: no Topical: yes	euphonium	Semantic: yes Topical: yes

Word2Vec AP Calculation

car

#	Closest Words-car							
	word2vec bow5	rel	prec	Σ	***word2vec: dependency***	rel	prec	Σ
1	cars	1	1	1	truck	1	1	1
2	truck	1	1	2	suv	1	1	2
3	automobile	1	1	3	vehicle	1	1	3
4	vehicle	1	1	4	minivan	1	1	4
5	motorbike	1	1	5	cars	1	1	5
6	motorcycle	1	1	6	speedboat	1	1	6
7	driver	1	1	7	racecar	1	1	7
8	minivan	1	1	8	automobile	1	1	8
9	suv	1	1	9	motorcar	1	1	9
10	lorry	0	9/10	9	jeep	1	1	10
11	motorcar	1	10/11	9.91	limousine	1	1	11
12	mid-engined	1	11/12	10.83	minibus	1	1	12
13	limousine	1	12/13	11.75	lorry	0	12/13	12
14	front-engined	1	13/14	12.68	limo	1	13/14	12.93
15	moped	1	14/15	13.61	motorcycle	1	14/15	13.86
16	motorhome	1	15/16	14.55	bike	1	15/16	14.8
17	mercedes-benz	1	16/17	15.49	motorhome	1	16/17	15.74
18	bike	1	17/18	16.43	taxicab	1	17/18	16.69
19	rear-engined	1	18/19	17.38	roadster	1	18/19	17.63
20	three-wheeled	1	19/20	18.33	wagon	1	19/20	18.58
N= 28 (union of all relevant words, from word2vec results only)		AP = 0.654			AP = 0.664			

piano

#	Closest Words-piano							
	word2vec bow5	rel	prec	Σ	word2vec: dependency	rel	prec	Σ
1	violin	1	1	1	violin	1	1	1
2	cello	1	1	2	cello	1	1	2
3	harpsichord	1	1	3	harpsichord	1	1	3
4	clarinet	1	1	4	saxophone	1	1	4
5	viola	1	1	5	clarinet	1	1	5
6	flute	1	1	6	guitar	1	1	6
7	bassoon	1	1	7	trombone	1	1	7
8	violoncello	1	1	8	mandolin	1	1	8
9	oboe	1	1	9	vibraphone	1	1	9
10	concerto	1	1	10	marimba	1	1	10
11	saxophone	1	1	11	accordion	1	1	11
12	accordion	1	1	12	pianoforte	1	1	12
13	harp	1	1	13	bassoon	1	1	13
14	trombone	1	1	14	fortepiano	1	1	14
15	sonatas	1	1	15	violoncello	1	1	15
16	trumpet	1	1	16	trumpet	1	1	16
17	mandolin	1	1	17	harmonica	1	1	17
18	pianoforte	1	1	18	clavinet	1	1	18
19	vibraphone	1	1	19	clavichord	1	1	19
20	concertos	1	1	20	euphonium	1	1	20
N= 27 (union of all relevant words, from word2vec results only)		AP = 0.741			AP = 0.741			