Context Similarity and Semantic Relationships

Nami Akazawa, Emre Canbazer
Supervised by Sylvain Pogodalla

MSc Natural Language Processing
University of Lorraine

Institut des sciences du Digital Management & Cognition

Loria – Laboratoire lorrain de recherche en informatique et ses applications

Abstract

Vector representations of words (vector space models) can capture different relationships exist in words by using context information embedded in the text. However, how linguistic processing impacts encoding of semantic relatedness in VSM is still unknown. We propose a dependency-based construction of VSM utilizing syntactic relations using MANGOES software and analyze the result.

Background

- The traditional word-based co-occurrence models build their vector space by only considering a window of co-occurring words surrounding the target word.

Sentences:

Simple strawberry pie with fresh strawberries coated in a light strawberry glaze.
My favorite pie is cherry pie but I like apple pie as well.
A doctor opens the medicine cabinet to get drugs.
I go pharmacy to get medicine that I need.

Window size: 4
Target words: ⟨*strawberry, cherry, doctor, pharmacy*⟩
Basis elements: ⟨*simple, strawberry, pie, ..., my, favorite, ..., a, doctor, opens, ... go, pharmacy, that, need*⟩

	simple	...	pie	favorite	medicine	drugs
strawberry	1	...	1	0	0	0
cherry	0	...	2	1	0	0
doctor	0	...	0	0	1	0
pharmacy	0	...	0	0	1	0

Figure 1: A simple example of co-occurring space given a set of sentences from a text, showing five of the dimensions (for pedagogical purpose).

- **Dependency-based VSM**
- The intuition of the syntax-based model is that we may construct a semantically-enriched word vector space model that captures different semantic relations by incorporating information about the syntactic relationship between a target word and other words.

Sentence:

He ate the cheese sandwich

Target words: ⟨*he, ate, cheese, sandwich*⟩
Basis elements: ⟨*(subj, he), (root, ate), (det, the), (mod, cheese), (obj, sandwich)*⟩

	(subj, he)	(root, ate)	(det, the)	(mod, cheese)	(obj, sandwich)
he	0	0	0	0	0
ate	1	0	0	0	1
cheese	0	0	0	0	0
sandwich	0	0	1	1	0

Figure 2: A simple example of Lin's (Lin, 1998) dependency-based semantic space.

- **Lexical Functions**
- Linguistic notion that explains the semantic relation between semantically related wordforms.
  - Paradigmatic LF's
    - **S1**(LECTURE) → LECTURER
    - **V0**(STYLE) → STYLIZE
    - **Syn**$_\cap$(PARTNER) → SIGNIFICANT OTHER
    - **Anti**(VICTORY) → DEFEAT
  - Syntagmatic LF's
    - **Magn**(RAIN) → HEAVY
    - **Oper1**(CRIME) → COMMIT
    - **FinOper1**(POWER) → LOSE
    - **Func0**(SNOW) → FALL

Corpora

Wikipedia extracted texts parsed using Stanford CoreNLP which outputs grammatical relations in the Universal Dependencies v1 representation in CoNLL-U format (includes lemma, POS, dependency relation and head word).

Language	English	French
Number of Sentences	19182562	18916629
Number of Tokens	518854512	494582365
Number of Unique Tokens	4793040	3949618
Number of Tokens without stop-words	4787871	3941659

Experiments

We used MANGOES software developed by by magnet team in INRIA. In total we built 33 models with different parameter settings. Some of the shared settings are:

- Max sentence length: 100
- Positive Point-wise mutual information as weighting function
- Removal of stop-words from target and context vocabularies
- Considers entity token in the form of (lemma, POS)

As our vocabulary setting, we have tried three different ways: applying POS filter to only target vocabulary, only context vocabulary, and both vocabulary. For each option, we applied (ADV,ADJ, NOUN, VERB), (NOUN, VERB, ADJ), (NOUN, ADV, VERB) respectively.

One can define how far a target word wants to include as its context using dependency connection. We tested paths of length 1 and 2 as context (depth) for the depth setting of dependency context. The depth of 2 considers path length of at most two.

We constructed three path value functions.

- **BASE** assigns the value of 1 to all counted paths. It assumes that all paths are equally important.
- ***Length*** assigns each path a value inversely proportional to its length. It discourages the weight to longer paths.
- ***Gram-rel*** defines ranking paths according to its dependency relations.

	(australian,ADJ)	(scientist,NOUN)	(discovers,VERB)	(star, NOUN)	(with, ADP)	(telescope,NOUN)
australian	0	1	3	0	0	0
scientist	1	0	3	3	0	3
discovers	3	3	0	1	2	2
star	0	3	1	0	0	2
with	0	0	2	0	0	1
telescope	0	3	2	2	1	0

Example of dependency-based co-occurrence matrix with weighted scheme of {nsubj : 3, nmod : 2} and others are map to 1

Settings for the VSMs shown in **Result** are following:

- **BASE**: depth 1 +base path value function
- **SVD2**: depth 2 +length path value function
- **SVD3**: English - depth 2 +gram-rel path value function with weight scheme of {pobj: 5, dobj: 5, iobj: 5,obj: 5, nsubj: 5, obl: 5}.
- **SVD4**: depth 2 + dependency relation filter of {pobj, dobj, iobj, obj, nsubj, obl}.
- **SVD5**: depth 2 + POS filter of {ADV, ADJ, NOUN, VERB} to both target and context vocabulary

Results

We performed a qualitative evaluation by collecting randomly chosen target words' top 5 similar words computed from VSM using cosine similarities and display the 5 word embeddings and their results. We manually analyze and discuss the outputs.

For English:

WORD	BASE	SVD2	SVD3	SVD4	SVD5
pain	cough, afflict, debilitate, suffer, bruise	painful, chronic, acute, severe, traumatic	painful, cough, fatal, traumatic, chronic	sick, ill, insane, painful, sudden	painful, cough, chronic, sweat, sore
price	non-monopoly, supra-competitive, re-roll, minimum, reasonable	worth, pay, net, cost, exceed	worth, net, cost, pay, financial	million, billion, chronically, multi-day, cost	worth, net, pay, profitable, exceed

Pain

*- Compared to the BASE, the other models describe the type of pain: acute pain and chronic pain which intensify the pain thus falls in category of **Magn**.*

*- SVD2 contains 'severe', which intensify the meaning of pain, can also be categorized as **Magn**.*

Price

*- BASE proposes rare word: non-monopoly and supra-competitive. SVD2, SVD3 and SVD5 captured the **Oper1** collocate of the target word: pay.*

For French:

WORD	Base setting	SVD2	SVD3	SVD4	SVD5
accord	laccord, pacte, conclure, prévoire, lacte	conclure, prévoir, accepté, décidé, ...	signé, conclure, signer, négocié, négocier	conclure, prévoir, signer, décider, accepter	signé, conclure, négocier, signer, prévoyant
argument	largument, priori, évidemment, ceci, suppose	saurait, clairement, justement, évident, justifier	évident, logique, savoir, justifier, contredire	évident, justement, clairement, saurait, logique	largument, logique, évident, contredire, savoir

Accord

- *SVD3 and SVD5 captures collocations like négocier* ('to negociate') and *signer* ('to sign'), which could be explained by the lexical functions ***IncepOper1*** and ***Caus1Func0*** .

Argument

- The collocate *logique* ('logical') captured by the models SVD3, SVD4 and SVD5 could be explained by ***Ver*** .

Future Directions

- More corpus analysis and pre-processing.
- Increase of corpus size.
- Further depth setting.
- Combination of parameters.