Vector Space Models of Meaning

LT2213/LT2813 V23: Computational Semantics

Nikolai Ilinykh and many others . . .
Department of Philosophy, Linguistics and Theory of Science
Centre for Linguistic Theory and Studies in Probability (CLASP)
University of Gothenburg, Sweden
nikolai.ilinykh@gu.se

Presented at, April 11, 2023

Where we are at

GÖTEBORGS UNIVERSITET

- we have looked at how to interpret and model the meaning of words and sentences with representations from logic
- now we will look at how to represent the meaning of words via **vector spaces** (we are not going to achieve that with logic-based representations)
- what we are covering today:

Where we are at

GÖTEBORGS UNIVERSITET

- we have looked at how to interpret and model the meaning of words and sentences with representations from logic
- now we will look at how to represent the meaning of words via **vector spaces** (we are not going to achieve that with logic-based representations)
- what we are covering today:
  - distributional (vector-space) semantics
  - measuring the semantic similarity of words in terms of the similarity of the contexts
  - application of vector spaces and word vectors for a downstream task

From similarity in nature . . .

(a) Smilodon

(b) Thylacosmilus

- mammals that exist(ed) far from each other might be pretty similar because of the environment that has led to their evolution
- this is one of the examples where **context** is important since it makes two completely distinct entities very similar with each other

. . . to similarity in language

- "The bear ran away"
- "The pear ran away"

. . . to similarity in language

- "The bear ran away"
- "The pear ran away"
- Similar meaning: airplane and aircraft, similar context: airplane and airline
- The type of similarity relationship varies between situations

. . . to similarity in language

- "The bear ran away"
- "The pear ran away"
- Similar meaning: airplane and aircraft, similar context: airplane and airline
- The type of similarity relationship varies between situations
- But **how much meaning we can capture from text alone?**

Learning meaning of words from context

- representing words through context is the example of **representation learning**
- learning meaning of words automatically, in **self-supervised** fashion, instead of conducting a lot of **feature engineering**
- we could identify the meaning of the word w based on its context (the set of words within a fixed-size window)
- most recent example of feature engineering: first assignment in the course, recognise yourself?

Thesaurus-based approaches are hard

- we do not have a thesaurus for every language
- thesauri have problems with recall
  - many words and phrases are missing
  - thesauri do not work good enough for verbs and adjectives
  - dynamic nature of language is the problem [1]

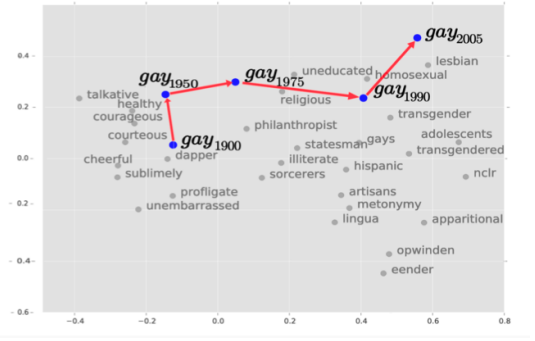
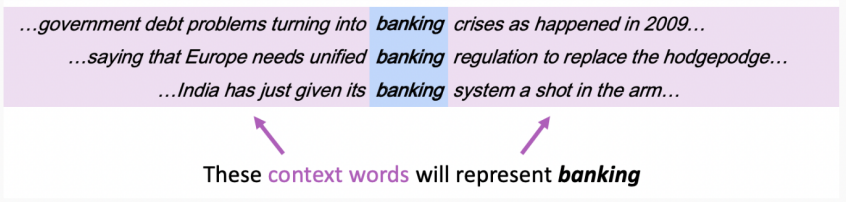


Figure: Kulkarni et. al 2014

[1] https://gupea.ub.gu.se/handle/2077/74969

Distributional Semantics

- **Distributional Hypothesis**: a word's meaning is defined by the words that frequently appear close-by
  - "You shall know a word by the company it keeps (Firth 1957)"
  - one of the core ideas of modern statistical NLP
- we could identify the meaning of the word *w* based on its context (the set of words within a fixed-size window): **word co-occurrence vector**



Words are hard! (@_@)

GÖTEBORGS UNIVERSITET

- MOUSE = “mouse”

Words are hard! (@_@)

GÖTEBORGS UNIVERSITET

- MOUSE = "mouse"
- ambiguity in meaning: what type of mouse? (*word sense disambiguation*)
- synonymy: mouse and rat; even better - couch and sofa

Words are hard! (@_@)

GÖTEBORGS UNIVERSITET

- MOUSE = "mouse"
- ambiguity in meaning: what type of mouse? (*word sense disambiguation*)
- synonymy: mouse and rat; even better - couch and sofa
- principle of contrast: when do we use "couch" or "sofa"? "water" or "H2O"?

Words are hard! (@_@)

GÖTEBORGS UNIVERSITET

- MOUSE = "mouse"
- ambiguity in meaning: what type of mouse? (*word sense disambiguation*)
- synonymy: mouse and rat; even better - couch and sofa
- principle of contrast: when do we use "couch" or "sofa"? "water" or "H2O"?
- word similarity: cat is not a synonym to dog, but cats and dogs are similar words (both are animals, mammals)
- word relatedness: words are often related *eventively*: coffee and cup, food and chef
- semantic frames and roles: X bought the book from Y = Y sold the book to X
- connotation: copy, replica, plagiarism, reproduction - are they the same?

Relations between words are useful

GÖTEBORGS UNIVERSITET

- Grammatical relations: subject-verb, verb-object, etc
  - edible items can be the direct object of verbs like eat, cook, serve, but not of other verbs such as drive
  - use parsers to extract syntactic relationships
- Non-grammatical relations: n-grams within a certain distance (window)
  - n-grams

Relations between words are useful

GÖTEBORGS UNIVERSITET

- Grammatical relations: subject-verb, verb-object, etc
  - edible items can be the direct object of verbs like eat, cook, serve, but not of other verbs such as drive
  - use parsers to extract syntactic relationships
- Non-grammatical relations: n-grams within a certain distance (window)
  - n-grams
- **A very important step** when designing a word similarity system is to select the type of co-occurrence between the words (e.g., grammatical, n-gram).

Roots of Vector Semantics (II)

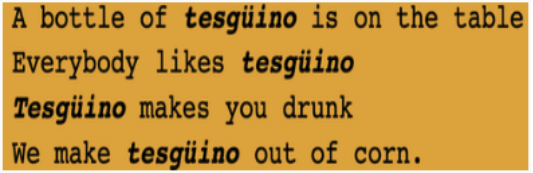


Figure: Joos 1950, Harris 1954, Firth 1957

- two words are similar if they have similar word contexts, e.g. tesguino means an alcoholic beverage like a beer
- **it means that we can use neighbouring words to represent our target word as a vector in multi-dimensional space**

What are the ways to represent words as vectors?

GÖTEBORGS UNIVERSITET

- sparse vector representations: word co-occurence matrices or mutual information representation
- dense vector representations: vectors learned with singular value decomposition (or latent semantic analysis), neural-network-inspired models (skip-grams, CBOW), Brown clusters

What are the ways to represent words as vectors?

GÖTEBORGS UNIVERSITET

- sparse vector representations: **word co-occurence matrices or mutual information representation**
- dense vector representations: vectors learned with **singular value decomposition** (or latent semantic analysis), neural-network-inspired models (skip-grams, CBOW), Brown clusters

General intuition

GÖTEBORGS UNIVERSITET

- we model the word meaning by "embedding" it in a vector space
- the meaning of a word is vector of numbers (such vector models are also called **embeddings**)
- contrast: in many NLP applications, the word meaning is represented by a vocabulary index ("word number 354")

General intuition

GÖTEBORGS UNIVERSITET

- we model the word meaning by "embedding" it in a vector space
- the meaning of a word is vector of numbers (such vector models are also called **embeddings**)
- contrast: in many NLP applications, the word meaning is represented by a vocabulary index ("word number 354")
- a funny joke (I think so):
  - Q: What is the meaning of life?
  - A: LIFE

Documents as contexts

GÖTEBORGS UNIVERSITET

- Let's say we have a corpus of texts and it consists of a (large) number of documents (articles, plays, novels, etc.)
- In that case, we can define the contexts of a word as the sets of documents in which it appears

Documents as contexts

GÖTEBORGS UNIVERSITET

- Let's say we have a corpus of texts and it consists of a (large) number of documents (articles, plays, novels, etc.)
- In that case, we can define the contexts of a word as the sets of documents in which it appears
- Conversely, we can represent each document as the (multi)set of words which appear in it
  - Intuition: documents are similar to each other if they contain the same words
  - This is useful for information retrieval, e.g. to compute the similarity between a query (a document) and any document in the collection to be searched

Term-Document Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

A term-document matrix is a 2D table:

- Each cell contains the frequency (count) of the term (word) in the document
- Each column is a vector of counts over words, representing a document
- Each row is a vector of counts over documents, representing a word

Term-Document Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

- Two documents are similar if their vectors are similar
- Two words are similar if their vectors are similar

Wrapping up the notion of context

GÖTEBORGS UNIVERSITET

- There are many different definitions of context that would produce different kinds of similarities:
- Context defined by nearby words
  - how often do words appear near the target word?
  - context words should occur frequently enough in our corpus to get reliable co-occurrence counts, but what to do with words which are too common?
  - define what is "nearby": **context window** is +- 5 words from the target word
  - get co-occurrence counts of words and contexts
  - define how to transform co-occurrence counts of words and contexts into vector element, e.g. PMI, SVD
  - define how to compute the similarity of word vectors, e.g. cosine

Wrapping up the notion of context

GÖTEBORGS UNIVERSITET

- There are many different definitions of context that would produce different kinds of similarities:
- Context defined by nearby words
  - how often do words appear near the target word?
  - context words should occur frequently enough in our corpus to get reliable co-occurrence counts, but what to do with words which are too common?
  - define what is "nearby": **context window** is +- 5 words from the target word
  - get co-occurrence counts of words and contexts
  - define how to transform co-occurrence counts of words and contexts into vector element, e.g. PMI, SVD
  - define how to compute the similarity of word vectors, e.g. cosine
- Contexts defined by grammatical relations
  - how often is specific noun used as the subject of the particular verb?
  - requires a parser
  - gives more fine-grained similarities

The notion of context

GÖTEBORGS UNIVERSITET

- There are many different definitions of context that would produce different kinds of similarities:
- **Context defined by nearby words**
  - how often do words appear near the target word?
  - context words should occur frequently enough in our corpus to get reliable co-occurrence counts, but what to do with words which are too common?
  - define what is "nearby": **context window** is +- 5 words from the target word
  - get co-occurrence counts of words and contexts
  - define how to transform co-occurrence counts of words and contexts into vector element, e.g. PMI, SVD
  - define how to compute the similarity of word vectors, e.g. cosine
- Contexts define by grammatical relations
  - how often is specific noun used as the subject of the particular verb?
  - requires a parser
  - gives more fine-grained similarities

Word-Word Matrix

We can get co-occurrence counts as either a binary feature or as a frequency count. Let's go with frequency counts.

sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer.**	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

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

Typically, we get 10k - 100k dimensions (contexts), vectors are very sparse.

Also, context window size would impact your representations: smaller window produces more syntactic representations, while bigger window tends to make your representation more semantic.

Similarity measures

GÖTEBORGS UNIVERSITET

- A mathematical measure for the degree of similarity between two pairs of vectors, we want to capture *association* between words

Similarity measures

GÖTEBORGS UNIVERSITET

- A mathematical measure for the degree of similarity between two pairs of vectors, we want to capture *association* between words
- $= freq(w_1, w_2)$
  - simple, but too crude; it over-emphasises the effect of words that are frequent in the corpus
  - a quick fix: normalise for word frequencies and possibly scale the association value according to individual weights of specific words

Similarity measures

GÖTEBORGS UNIVERSITET

- A mathematical measure for the degree of similarity between two pairs of vectors, we want to capture *association* between words
- $= freq(w_1, w_2)$
  - simple, but too crude; it over-emphasises the effect of words that are frequent in the corpus
  - a quick fix: normalise for word frequencies and possibly scale the association value according to individual weights of specific words
- $= association(w_1, attr)$ vs $association(w_2, attr)$
  - more sophisticated than simple frequencies

Using word vectors [2]

- vectors help us to build models that can learn differences and similarities between words
- let's say, we have the following vector space:

	animal	stable	village	gallop	jokey
horse	0	6	2	10	4
run	1	8	4	4	0

[2] http://semdial.org/anthology/Z22-Dobnik_semdial_0017.pdf

Using word vectors [2]

- vectors help us to build models that can learn differences and similarities between words
- let's say, we have the following vector space:

	animal	stable	village	gallop	jokey
horse	0	6	2	10	4
run	1	8	4	4	0

Short break time =)

[2] http://semdial.org/anthology/Z22-Dobnik_semdial_0017.pdf