Informatics 1 Cognitive Science

Lecture 12: Vector Semantics

Frank Keller

9 February 2024

School of Informatics
University of Edinburgh
keller@inf.ed.ac.uk

Slide credits: Frank Mollica, Chris Lucas, Mirella Lapata

Overview

Recap: Word Learning

In the last lectures, we discussed word learning and communicative efficiency:

- Word learning is helped by inductive biases and learning heuristics (fast mapping; cross-situational learning).
- Languages and language processing are optimized for effective communication; notions from information theory are useful to capture this.

In this lecture:

- We'll look at word meaning beyond reference to objects.
- We'll see how meaning representations can be learned from context.
- We'll test if these representations are cognitively plausible.
- Neural networks will make a comeback.

Time for a short quiz on Wooclap!

https://app.wooclap.com/GNJYGP

Meaning as Context

Context Vectors

We have seen in lecture 10 that reference is an important aspect of the meaning. However, what about about words that don't have an obvious referent:

- abstract nouns, verbs, adverbs, adjectives
- function words such as *every*, *but*, *from*, *they*

Another important component of meaning is the context in which a word is used. This idea goes back to philosophers such as Wittgenstein.

Context Vectors

We have seen in lecture 10 that reference is an important aspect of the meaning. However, what about about words that don't have an obvious referent:

- abstract nouns, verbs, adverbs, adjectives
- function words such as *every*, *but*, *from*, *they*

Another important component of meaning is the context in which a word is used. This idea goes back to philosophers such as Wittgenstein.

> *"She filled the samovar with water and heated it up to make tea."*

Context Vectors

We have seen in lecture 10 that reference is an important aspect of the meaning. However, what about about words that don't have an obvious referent:

- abstract nouns, verbs, adverbs, adjectives
- function words such as *every*, *but*, *from*, *they*

Another important component of meaning is the context in which a word is used. This idea goes back to philosophers such as Wittgenstein.

> *"She filled the samovar with water and heated it up to make tea."*

If you don't know the meaning of *samovar*, you can guess it from context.

Context Vectors

We have seen in lecture 10 that reference is an important aspect of the meaning. However, what about about words that don't have an obvious referent:

- abstract nouns, verbs, adverbs, adjectives
- function words such as *every*, *but*, *from*, *they*

Another important component of meaning is the context in which a word is used. This idea goes back to philosophers such as Wittgenstein.

> *"She filled the samovar with water and heated it up to make tea."*

If you don't know the meaning of *samovar*, you can guess it from context.

We can use this idea to learn meaning representations called word vectors or word embeddings.

Context Vectors

We construct context vectors using a window around the words we're interested in (target words):

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later ...

Context Vectors

We construct context vectors using a window around the words we're interested in (target words):

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later ...

Context Vectors

We construct context vectors using a window around the words we're interested in (target words):

. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later . . .

Context Vectors

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their $\boxed{\textit{symbols. these must first be examined in}}$ *the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his* $\boxed{\textit{initial understanding. to learn the meaning of}}$ *symbols is part of the anthropologist's* $\boxed{\textit{practical semantics: to discover the meaning of}}$ *words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts.* $\boxed{\textit{these must come first; fantasy can come later}}$ *...*

Context Vectors

. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later . . .

first

learn

discover

Context Vectors

. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their $\boxed{\textit{symbols. these must first be examined in}}$ *the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his* $\boxed{\textit{initial understanding. to learn the meaning of}}$ *symbols is part of the anthropologist's* $\boxed{\textit{practical semantics: to discover the meaning of}}$ *words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts.* $\boxed{\textit{these must come first; fantasy can come later}}$ *. . .*

	these	meaning	to	practical	come
first					
learn					
discover					

Context Vectors

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later ...

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Context Vectors

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Context Vectors

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

Context Vectors

Context words

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

Context Vectors

$$
\begin{array}{l}
\\
\\
\text{first} \\
\text{learn} \\
\text{discover}
\end{array}
$$

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Context words (over the column headers)

Target words (under the row labels) — Context vectors (under the values)

Context Vectors

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Context words

Target words

Context vectors

Informal algorithm for constructing context vectors:

- pick the words you are interested in: target words;
- define number of words around target word: context window;
- count number of times the target word co-occurs with each context word: context vector.

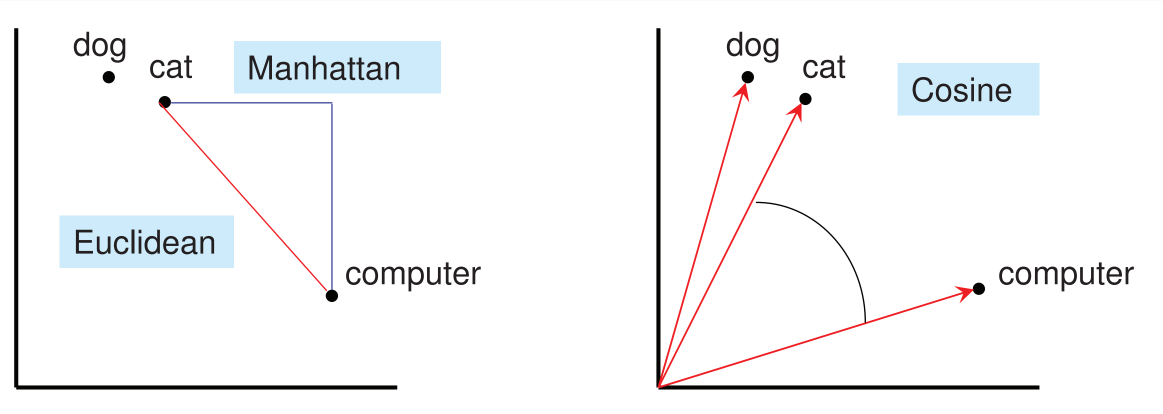
Time for a short quiz on Wooclap!

https://app.wooclap.com/GNJYGP

Measuring Similarity

Comparing Context Vectors

Measure the distance between vectors:



Measures of Distributional Similarity

The cosine of the angle between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is:

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

The Euclidean distance of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is:

$$||\boldsymbol{x} - \boldsymbol{y}|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Many more similarity measures exist.

Context Vectors for Word Meaning

- Context vectors can be used to learn syntactic categories.
- Syntactic category learning typically uses a small context window (e.g., 3 words);
- If we use a larger context window (could be the whole document), then context vectors capture *word meaning* well;
- Many examples of this approach; most prominent one is *Latent Semantic Analysis* (LSA).
- Context vectors are normally too sparse (too many zeros).
- LSA therefore contains a *dimensionality reduction step:* essentially, merge dimensions that are similar across vectors.

Representation Learning

Learning Context Vectors

Counting vs. learning:

- Traditional approach: count word co-occurrences and compress the resulting sparse vectors by dimensionality reduction;
- this is used in traditional models like LSA;
- But what if we could *learn* good context vectors rather than just getting them by *counting?*

Learning Context Vectors

Counting vs. learning:

- Traditional approach: count word co-occurrences and compress the resulting sparse vectors by dimensionality reduction;
- this is used in traditional models like LSA;
- But what if we could *learn* good context vectors rather than just getting them by *counting?*

This is the idea behind *word embeddings.* These are essentially clever context vectors learned using neural networks.

Learning Context Vectors

Counting vs. learning:

- Traditional approach: count word co-occurrences and compress the resulting sparse vectors by dimensionality reduction;
- this is used in traditional models like LSA;
- But what if we could *learn* good context vectors rather than just getting them by *counting?*

This is the idea behind *word embeddings.* These are essentially clever context vectors learned using neural networks.

But how could we get a neural net to learn context vectors?

Learning Context Vectors

Key idea: *train a neural network to guess a word from its context:*

- input: representation of the context;
- output: representation of the word;
- training data: words within a context window.

This model is based on same information as count vectors (the word and a window of context words), but uses it differently.

Learning Context Vectors

Key idea: *train a neural network to guess a word from its context:*

- input: representation of the context;
- output: representation of the word;
- training data: words within a context window.

This model is based on same information as count vectors (the word and a window of context words), but uses it differently.

It is an example of using self-training for neural networks.

Learning Context Vectors

Key idea: *train a neural network to guess a word from its context:*

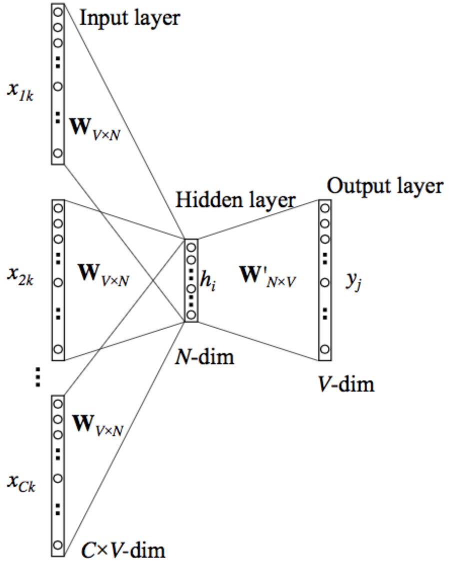
- input: representation of the context;
- output: representation of the word;
- training data: words within a context window.

This model is based on same information as count vectors (the word and a window of context words), but uses it differently.

It is an example of using self-training for neural networks.

But if we do this, how do we get our context vectors?

word2vec Model



word2vec Model

- The word2vec model (Mikolov et al. 2013) uses context words to predict the target word;
- a neural net with only a single, linear hidden layer is used;
- the weights for different input positions are shared;
- the context window is of size five (two context words each to the left and right of the target word);
- no representation of word order: all context words are treated in the same way ("bag of words").

Representation Learning

The input and output layers are *one-hot encoded:*

Each word is represented as a vector of size V (the number of words in the vocabulary), where one unit is 1, and all others are 0.

```
             Paris
        Rome                          word V
Rome   = [1, 0, 0, 0, 0, 0, ..., 0]
Paris  = [0, 1, 0, 0, 0, 0, ..., 0]
Italy  = [0, 0, 1, 0, 0, 0, ..., 0]
France = [0, 0, 0, 1, 0, 0, ..., 0]
```

Representation Learning

The training examples are target words and their contexts:

initial understanding. to learn the meaning of symbols is part of the anthropologist's practical $\boxed{\textit{semantics: to discover the meaning}}$ *of words, noticing when their use is appropriate and when it is not. all this requires*

For this example, the input is: $x_1 =$ semantics, $x_2 =$ to, $x_3 =$ the, $x_4 =$ meaning. And the output is $y_i =$ discover.

The model is trained on lots of examples like this using backpropagation.

Representation Learning

The training examples are target words and their contexts:

initial understanding. to learn the meaning of symbols is part of the anthropologist's practical *semantics: to discover the meaning* *of words, noticing when their use is appropriate and when it is not. all this requires*

For this example, the input is: $x_1 =$ semantics, $x_2 =$ to, $x_3 =$ the, $x_4 =$ meaning. And the output is $y_i =$ discover.

The model is trained on lots of examples like this using backpropagation.

After training, the *hidden layer* h_i for target word y_i can be used as its *context vector* (also called a word embedding).

Model Evaluation

What can we do with these word embeddings? *We can answer questions about words!*

Task: for a question word, find an answer word that's syntactially or semantically related. You are given a list of possible answer words.

Solution: take the word embedding for the question word and compare it to the word embeddings for the answer words. Return the one that's most similar (use cosine as similarity measure).

Training: 6 billion words of text; 1M word vocabulary.
Testing: 20k questions–answer pairs.

Model Evaluation

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Model Evaluation

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

word2vec gets up to 55% accuracy for the semantic relationships, and up to 64% for the syntactic relationships (Mikolov et al. 2013).

Model Evaluation

We can subtract two context vectors, then add the result to another context vector:

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

Check more examples out at:

http://rare-technologies.com/word2vec-tutorial/

Modeling Semantic Priming

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole.**

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole.**

ground

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole.**

ground face

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole.**

ground face drown

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard. Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole.**

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole.**

ground

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole.**

ground face

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole.**

ground face drown

Semantic Priming

These results are impressive, but can we use word embeddings to model *actual cognitive data?*

Let's look at a *lexical decision experiment* by Till, Mross and Kintsch's (1988). Participants saw stimuli such as:

> The gardener pulled the hose around to the holes in the yard.
> Perhaps the water would solve his problem with the **mole.**

ground face drown cancer

> The patient sensed that this was not a routine visit.
> The doctor hinted that there was reason to remove the **mole.**

ground face drown cancer

Simulating Semantic Priming

Till, Mross and Kintsch's (1988) results:

- words related to both senses of the ambiguous word were primed immediately after presentation;
- after about 300 ms only the context appropriate associates remained significantly primed.

Word embeddings predict:

- larger cosines between ambiguous word and related word compared to control word;
- vector average of the context words has a higher cosine with semantically congruent words.

Simulating Semantic Priming

The patient sensed that this was not a routine visit.
The doctor hinted that there was reason to remove the mole.

ground	face	drown	cancer
.15	.24	.15	.21

The TOEFL Task

Test of English as a Foreign Language tests non-native speakers' knowledge of English. Part of the test is a *synonym task:*

> You will find the office at the main intersection.
>
> (a) place
> (b) crossroads
> (c) roundabout
> (d) building

This is a standard task in the cognitive modeling literature, and context vectors are frequently used to solve it.

The TOEFL Task

Test of English as a Foreign Language tests non-native speakers' knowledge of English. Part of the test is a *synonym task:*

> You will find the office at the main intersection.
>
> (a) place
>
> (b) crossroads
>
> (c) roundabout
>
> (d) building

This is a standard task in the cognitive modeling literature, and context vectors are frequently used to solve it.

The TOEFL Task

Use word2vec trained the Google News dataset (100 billion words). The resulting word embeddings are 300 dimensional.

- The TOEFL dataset has 80 items: 1 word/4 alternative words.
- Compute word embeddings for probe and answer words.
- Word with largest cosine to the probe is correct answer.
- word2vec answered around 80% of items correctly.
- Non-native speakers' average is 64.5%.

Pereira et al. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. Cog. Neuropsychology. http://dx.doi.org/10.1080/02643294.2016.1176907

Possible Problems with Word Vectors

Time for a short quiz on Wooclap!

https://app.wooclap.com/GNJYGP

Discussion

Strengths:

- learns word representations automatically from raw text;
- simple approach: all we need is a corpus and some notion of what counts as a word;
- language-independent, cognitively plausible.

Weaknesses:

- many ad-hoc parameters when creating the embeddings;
- ambiguous words: their meaning is the average of all senses;
- no representation of word order:

> The author received much acclaim for his new **book.**
> For author acclaim his much received new **book.**