Word Embeddings

CS 6956: Deep Learning for NLP

THE UNIVERSITY OF UTAH

Overview

- Representing meaning
- Word embeddings: Early work
- Word embeddings via language models
- Word2vec and Glove
- Evaluating embeddings
- Design choices and open questions

Overview

- Representing meaning
- Word embeddings: Early work
- Word embeddings via language models
- Word2vec and Glove
- Evaluating embeddings
- Design choices and open questions

Representing meaning

What do words mean?

How do they get their meaning?

Representing meaning

What do words mean?

How do they get their meaning?

tiger cat dog table

Representing meaning

What do words mean?

How do they get their meaning?

tiger

cat

dog

table

Representing meaning

What do words mean?

How do they get their meaning?

tiger cat dog table

Perhaps more pertinent for modeling language:

How can we represent the meaning of words in a form that is computationally flexible?

Words are atomic symbols

The strings `cat`, `tiger`, `dog` and `table` are different from each other

If we systematically replace all words with unique identifiers, does their meaning change?

- Think about substituting `cat` with `uniq-id-1`, `table` with `uniq-id-53`, …

- As long as we are consistent in our substitution, sentence meaning would not be harmed

Words are atomic symbols

The strings `cat`, `tiger`, `dog` and `table` are different from each other

If we systematically replace all words with unique identifiers, does their meaning change?

- Think about substituting `cat` with `uniq-id-1`, `table` with `uniq-id-53`, …

- As long as we are consistent in our substitution, sentence meaning would not be harmed

So how do we represent word meaning in a way that is grounded in the way they are used by everyone?

Words are atomic symbols

The strings `cat`, `tiger`, `dog` and `table` are different from each other

If we systematically replace all words with unique identifiers, does their meaning change?

- Think about substituting `cat` with `uniq-id-1`, `table` with `uniq-id-53`, …

- As long as we are consistent in our substitution, sentence meaning would not be harmed

So how do we represent word meaning in a way that is grounded in the way they are used by everyone?

Various perspectives exist

An ontology: Eg. WordNet

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun cat
8 senses of cat
Sense 1
cat, true cat
=> feline, felid
Sense 2
guy, cat, hombre, bozo
=> man, adult male
Sense 3
Cat
=> gossip, gossiper, gossipmonger, rumormonger, rumourmonger, newsmonger
Sense 4
kat, khat, qat, quat, cat, Arabian tea, African tea
=> stimulant, stimulant drug, excitant
Sense 5
cat-o'-nine-tails, cat
=> whip
Sense 6
Caterpillar, cat
=> tracked vehicle
Sense 7
big cat, cat
=> feline, felid
Sense 8
computerized tomography, computed tomography, CT, computerized axial tomography, computed axial tomography, CAT
=> X-raying, X-radiation

An ontology: Eg. WordNet

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun cat

8 senses of cat

Sense 1

cat, true cat

=> feline, felid

Such a taxonomy shows hypernymy relationships between words

Sense 2

guy, cat, hombre, bozo

=> man, adult male

Sense 3

Cat

=> gossip, gossiper, gossipmonger, rumormonger, rumourmonger, newsmonger

Sense 4

kat, khat, qat, quat, cat, Arabian tea, African tea

=> stimulant, stimulant drug, excitant

Sense 5

cat-o'-nine-tails, cat

=> whip

Sense 6

Caterpillar, cat

=> tracked vehicle

Sense 7

big cat, cat

=> feline, felid

Sense 8

computerized tomography, computed tomography, CT, computerized axial tomography, computed axial tomography, CAT

=> X-raying, X-radiation

An ontology: Eg. WordNet

Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun cat
8 senses of cat
Sense 1
cat, true ca
  => feline, felid
Sense 2
guy, cat, hombre, bozo

Such a taxonomy shows hypernymy relationships between words

- A high precision resource
- Typically manually built
  - Hard to keep it up-to-date
  - New words enter our lexicon, words change meaning over time
- Does not necessarily reflect how words are used in real life
  - Perhaps related to the previous concern
- Various methods for computing similarities between words using such an ontology.
  - Eg: using distances in the hypernym hierarchy such as the Wu & Palmer similarity measure

computerized tomography, computed tomography, CT, computerized axial tomography, computed axial tomography, CAT
  => X-raying, X-radiation

The distributional hypothesis

Words that occur in the same context have similar meanings

- Zelig Harris, J. R. Firth
- Firth (1957) : “You shall know a word by the company it keeps”

The distributional hypothesis

Words that occur in the same context have similar meanings

- Zelig Harris, J. R. Firth
- Firth (1957) : “You shall know a word by the company it keeps”

- The key idea: To characterize the meaning of a word, we need to we characterize the distribution of its context

The distributional hypothesis

Words that occur in the same context have similar meanings

- Zelig Harris, J. R. Firth
- Firth (1957) : “You shall know a word by the company it keeps”

- The key idea: To characterize the meaning of a word, we need to we characterize the distribution of its context

John sleeps during the		and works at night
Mary starts her	***day***	with a cup of coffee
He starts his		with an angry look at his inbox
⋮		⋮

The distributional hypothesis

Words that occur in the same context have similar meanings

- Zelig Harris, J. R. Firth
- Firth (1957) : “You shall know a word by the company it keeps”

- The key idea: To characterize the meaning of a word, we need to we characterize the distribution of its context

context

John sleeps during the		and works at night
Mary starts her	***day***	with a cup of coffee
He starts his		with an angry look at his inbox
⋮		⋮

The distributional hypothesis

Words that occur in the same context have similar meanings

- Zelig Harris, J. R. Firth
- Firth (1957) : "You shall know a word by the company it keeps"

- The key idea: To characterize the meaning of a word, we need to we characterize the distribution of its context

- What context?

  Commonly interpreted as neighboring words in text, but could be syntactic/semantic/discourse/pragmatic/… context.

  We will see more about context soon

Symbolic vs. Distributed representations

- The words `cat`, `tiger`, `dog` and `table` are symbols
- Just knowing the symbols does not tell us anything about what they mean. For example:
  1. Cats and tigers are conceptually closer to each other than to dogs or tables
  2. Cats, tigers and dogs are closer to each other than tables

Symbolic vs. Distributed representations

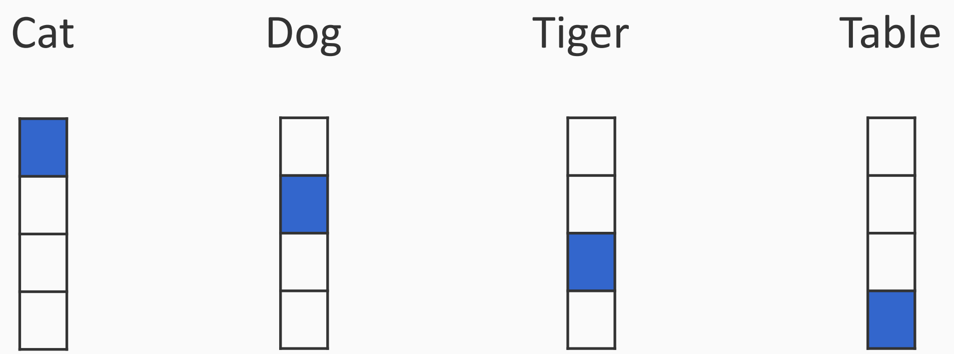
- The words `cat`, `tiger`, `dog` and `table` are symbols

- Just knowing the symbols does not tell us anything about what they mean. For example:
    1. Cats and tigers are conceptually closer to each other than to dogs or tables
    2. Cats, tigers and dogs are closer to each other than tables

- What we need: A representation scheme that inherently captures similarities between similar objects

Symbolic vs. Distributed representations

For example: Think about feature representations



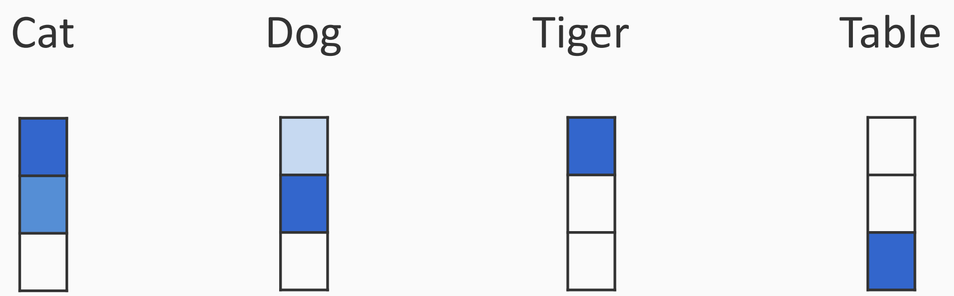
These ***one-hot vectors*** do not capture inherent similarities

Distances or dot products are all equal

Symbolic vs. Distributed representations

Distributed representations capture similarities better

– Think of them as vector valued representations can coalesce superficially distinct objects



Dense vector (often lower dimensional) representations can capture similarities better

Word embeddings (or word vectors)

A mapping from words to a vector space

- Could be a fixed mapping that is context independent (word2vec, Glove, etc)

  We will see these very soon

- Could be a parameterized mapping that is context dependent (ELMo, BERT, etc)

  We will see these later in the semester

A first step in any neural network model for textual inputs

- First, convert words to vectors, then attend to the task you want to solve

Perspectives on word embeddings

1. They capture distributional semantics: Embeddings are low dimensional vectors that are constructed by appealing to the distributional hypothesis

2. They are distributed representations of words: The embedding dimensions represent underlying aspects of meaning and words are characterized by membership to these latent dimensions

3. They provide features: Word embeddings are a widely-used, convenient ***learned*** feature representations.