Distributional Semantics

Lecture 1

Daniil Vodolazsky, Florian Gouret, Amir Bakarov
Novosibirsk State University, 2019

Introduction

What this course is about?

- Distributional Semantics is a science of how to represent the meaning of linguistic structures (particularly, words) through computational analysis of contexts of their use.
- In simple words, it is a science of learning a machine to understand the meaning of words (and, hence, texts).
- It is a highly interdisciplinary field that emerges linguistics, computer science, algebra, statistics, probability theory, psychology, etc
- Distributional semantics models are important tools used in a field of machine learning. Almost any modern natural language processing system (chatbots, recommender systems, information retrieval, etc) is based on distributional semantics.

What the aim of the course?

- To give an extensive overview of a field of distributional semantics.
- To look at distributional semantics from the perspective of linguistics as well as from perspective of computer science and maths and try understand how could they emerge.
- To understand how LDA, Word2Vec, FastText and other ubiquitous machine learning models work
- To give a broad picture of where distributional semantics is currently applied and where it could be successfully applied (may be you will be one who will do that).

Disclaimers

- This is not a natural language processing course. We will omit many important topics inside this field, focusing only on distributional semantics.
- We will have some lectures more related to linguistic side as well as some lectures related to computational side. As for the first part, you do not need any special knowledge about linguistics. However, you should be aware of some basics of linear algebra, statistics and probability theory. Basic Python coding skills are also required.
- This is a very pilot course, therefore some aspects would not be ideal. Your feedback and critics will help us to get better.

Syllabus

- Introduction
- Count-based Distributional Models
- Topic Modeling
- Prediction-based Distributional Models
- Lexical-level and Morphological-level Extensions of Word2Vec
- Evaluation of Distributional Semantic Models
- From Words to Phrases, Sentences and Documents
- Multi-sense Word Embeddings
- Cross-language Word Embeddings
- Recent Trends in Distributional Semantics
- Bridging Logics, Formal Semantics and Distributional Semantics
- Algebraic Semantics and Quantum Semantics
- Bridging Distributional Semantics and Neuroscience
- Conclusion

Organizers

- Daniil Vodolazsky (daniil.vodolazsky at mail.ru)
- Florian Gouret (fba.gouret at gmail.com)
- Amir Bakarov (amirbakarov at gmail.com)

How does it work

Meaning of Words

Semantics

- Semantics is a science that tries to define meaning of a sign (e.g. word)
- Lexical semantics is a science that tries to define the meaning of a word
- It tries to explain its flexibility in context and how it contributes to the construction of the meaning of sentences.
- Lexical semantics proposes different theories that try to explain this (as well as the nature of the concept “word” and the nature of concept “meaning”)
- Distributional semantics is one of such theories.

What does word “word” mean?

For the convenience in this course we will refer to word as a linguistic structure that requires to 3 following conditions:

1. A character set associated with it separates space characters from other words (we will omit languages that do not have a written tradition);
2. A phoneme set associated with it can be pronounced in separation from other phonemes;
3. It has a unitary meaning which could not be split into several meaning components.

Lexemes

- A stricter concept for analysis is a lexeme which is a set of word forms that have the same meaning.
- word forms: “cat”, “cats”. lexeme: “cat”.
- The lemma corresponds to the lexeme in a dictionary form (roughly speaking, the normal form of a word).

Different levels of the language

- Phonology
- Morphology
- Syntax
- Semantics
- Discourse/Pragmatics

Or (not a direct mapping!)

- Sounds
- Morphemes
- Words
- Sentences
- Texts

A notion of meaning

The theory of linguistic meaning has two related goals:

- to establish what words mean (this is the goal of lexical semantics)
- to explain how complex expressions acquire their meaning based on the meaning of their constituent parts (this is the purpose of a phrase or sentence semantics).

A notion of meaning

- Word = lexical information + encyclopedic information

  Lexical (embedded information):

1. Graphematic form
2. Phonology
3. Word class
4. Meaning

Encyclopedic information = knowledge about word

A notion of meaning

- Minimalist hypothesis: nothing that we know about word is embedded into its meaning.
- Maximalist hypothesis: all that we know about word is embedded in its meaning.

Theories of Lexical Semantics

Primary hypotheses

- Referential hypothesis
- Conceptual hypothesis
- Structural hypothesis
- Prototypical hypothesis
- Distributional hypothesis

Referential hypothesis

- Referential theory says that the words are used to refer on objects and events of the real word.
- Meaning is the ability to make a reference with the real world.
- One can describe each word meaning through the set of rules that create references through other words and objects.
- For example, by saying word `cat' we refer to a certain object with certain properties, and its extension includes all cats in the world.

Conceptual hypothesis

- Conceptual theory is based on the idea that the reference established between the word and the object is mediated by our mental representation of this object.
- The word obtain meaning only through this representation, and its called concept.
- When we say `cat', we are talking not about the real cat, but only about our metal representation of this object.

Structural hypothesis

- Structural hypothesis says that words meaning is not only limited by the ability to make a reference with the real world or some mental representation.
- The meaning of the word also includes a certain value, which is determined relatively to other words, particularly, the most similar words.
- So we could say that words meaning is defined by its semantic field.

Prototypical hypothesis

- Prototypical hypotheses is based on the notions of a category and a best exemplar of the category.
- Best exemplar is called prototype, and the meaning of the word is the set of the prototype's properties as well as set of objects in order for descending their closeness to the prototype.

Distributional hypothesis

- It is the idea that word meaning is determined by its contexts, or ‘You shall know a word by the company it keeps.’, as Firth, 1957 puts it.
- Words that share similar contexts should have similar meaning (e.g. “mug” and “cup”)
- We can represent meaning of the word by counting all context of its use.

Thank you