Chapter 1

Introduction

Keywords: lexicology, lexical semantics, lexicography, wordnet, distributional hypothesis, distributional semantics, vector space models.

1.1 What Is This Course About?

Distributional semantics is one of the most important notions in contemporary computational linguistics and NLP: representations of meaning exploiting distributional semantics framework are used in almost any NLP system. Therefore, deep understanding of distributional semantics and models based on it is crucial for resolving cutting-edge NLP tasks. The main idea of this course is to give such understanding to the listener.

It is important to understand that this course is not about NLP and/or semantics in general. Many aspects crucial for a NLP (or semantics) course will be omitted, and this course is not recommended for a listener seeking for a good introductory NLP course. Instead of it, this course will give an exhaustive (and unique) introduction to a much more narrower NLP field called distributional semantics (with a context of modern NLP techniques and linguistic theories). So, the course follows three main aims:

1. to give an understanding of distributional hypothesis and its role in contemporary language studies/technology;
2. to survey of state-of-the-art models and formalisms exploiting distributional hypothesis;
3. to describe application basis of distributional hypothesis.

It is also possible to consider this course as a very comprehensive overview of the field which also explains some basic NLP concepts. Despite this course proposes some very theoretical views on semantic theory, the main focus of the course would be held on computational and practical applications of distributional semantics. The course also requires a basic understanding of calculus, linear algebra and probability theory. No exceptional linguistic or engineering background is required.

To sump up: this course is about a narrow, but a very ubiquitous field of computational linguistics and NLP. It will give a deep understanding of what could be done with the most outstanding semantic theory of modern linguistics.

1.2 Basic Lexical Semantics

Lexical semantics is a science that tries to define the meaning of words, explaining its flexibility in context and how it contributes to the construction of the meaning of sentences. Lexical semantics is a subfield of lexicology, the science that studies words in the language (lexicon). One of the ultimate goals of lexicology is to develop a formal system representing the variety of lexicon.

Lexical semantics proposes different theories that try to explain the nature of word's meaning, – and, moreover, the nature of concept 'word' and the nature of concept 'meaning'. Distributional semantics is one of such theories, and it is important to understand the object of this study before going deeper.

1.2.1 A Notion of Word

The first thing essential for the concept 'word' is that this concept is very vague. Finding a formal definition refers to the philosophy of a language, and 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 (it is also a vague concept) which could not be split into several meaning components. From this perspective, we should not rely only on semantic criterion while separating words from non-words (since phrases like 'airbag' could be referred as word and not-word at the same time). We will exploit definition of unitary meaning (referred as a 'semantic component') of (Cruse 1986) to deal with this stumbling stone.

There are also different types of words, for example, Incorporated compounds, Juxtaposed compounds, etc. Fixed phrases. Words are divided according to the typology of languages (inflectional, isolating, etc.). At the same time, one word can have many forms (for simplicity, you can agree that all word forms are different words). We will omit different linguistic theories of words.

A stricter concept for analysis is a lexeme. Lexeme is a set of word forms that have the same meaning. The lemma corresponds to the lexeme in a dictionary form (roughly speaking, the normal form of a word).

1.2.2 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) and 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).

When we speak of the lexicon, we mean mental representation. Lexical information is embedded in the lexicon without representation (i.e., without definitions, as in a dictionary). The word is a lexical form, combined with concept and meaning.

What is included in the concept of 'value' is an open question relating to the prerogative of the philosophy of language. Nevertheless, we know for sure that the lexicon, like any

semiotic system, has two dimensions - the form and the content. Moreover, if there is content that does not have certain forms, then they can be viewed as concepts - mental categories with informational content, as if language-independent. Establishing the correspondence of a concept with a lexical form is called lexicalization - in the process of this a word is formed.

What distinguishes the meaning of a word from the meaning of syntactic and morphological units is that the meaning of a word can be perceived and described by the rapporteur more directly. That is, it is necessary to distinguish lexical meaning and grammatical meaning.

1.2.3 Theories of Lexical Semantics

We can distinguish 5 primary meaning theories that propose their own view on what is lexical semantics:

1. Referential hypothesis
2. Conceptual hypothesis
3. Structural hypothesis
4. Prototypical hypothesis
5. Distributional hypothesis

Referential hypothesis

Firstly introduced in 1982 (**?**).

Referential theory says that the words are used to refer on objects and events of the real word. So the **meaning of the word** is the ability to make a reference with the real world. This idea could be formalized by applying logical induction, and in formal semantics it is called **set theory**. In other words, we 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.

The referential hypothesis exploits the concept of intension, which means mapping of all possible worlds to their extensions. In other words, intension is a function which, given a word, returns the things denoted by that world in a particular world. It allows us to make sense of the fact that objects like 'my cat' and 'the pet that lives in my house' have different connotations despite they refer to the same objects.

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. In other words, the word obtain meaning only through this representation, and its called concept. And when we say 'cat', we are talking not about the real cat, but only about our metal representation of this object.

The framework that exploits conceptual hypothesis is called conceptual semantics, and it is ubiquitous among cognitivists and psycholinguists.

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 hypothesis 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.

This hypothesis exploits different interpretations of prototype, and it could figure as an analogue of a concept as well as some kind of core embedded into meaning and a word.

1.3 Computational Lexical Semantics

e.g. what tasks could be solved.

1.3.1 Historical Approaches

Levenshtein, dictionaries, etc.

1.3.2 Word Similarity

Thesauri, WordNet.

1.4 Distributional Hypothesis

Distributional hypothesis was first formulated by Harris, and by other linguists. 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. One can see that the words to the right and to the left of the words (their neighbors) in natural texts tend to be to some extent similar. These are the words like etc. It seems that semantically similar words share similar contexts.

1.5 Connection with Corpus Linguistics

Bibliography

[1] Geeraerts, D. (2010). *Theories of lexical semantics.* Oxford University Press.

[2] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

[3] Jezek, E. and Ježek, E. (2016). *The lexicon: An introduction.* Oxford University Press.

[4] Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.