Corpus Linguistics

Formal vs. Distributional Semantics

Anca Dinu, University of Bucharest,
February, 2013
Corpus Linguistics

• Corpus linguistics is a study of language and a method of linguistic analysis which uses a collection of natural or “real word” texts known as corpus.

• What Corpus Linguistics Does:
  – Gives an access to naturalistic linguistic information, “real word” texts which are mostly a product of real life situations. This makes corpora a valuable research source for grammar, semantics, dialectology, sociolinguistics, stylistics, etc.
Corpus Linguistics

– Facilitates linguistic research. Electronically readable corpora have dramatically reduced the time needed to find particular words or phrases. A research that would take days or even years to complete manually can be done in a matter of seconds with the highest degree of accuracy.

– Enables the study of wider patterns and collocation of words. Before the advent of computers, corpus linguistics was studying only single words and their frequency. Modern technology allowed the study of wider patterns and collocation of words.
Corpus Linguistics

– Allows analysis of multiple parameters at the same time. Various corpus linguistics software programmes and analytical tools allow the researchers to analyse a larger number of parameters simultaneously. In addition, many corpora are enriched with various linguistic information such as annotation.

– Facilitates the study of the second language. Study of the second language with the use of natural language allows the students to get a better “feeling” for the language and learn the language like it is used in real rather than “invented” situations.
Corpus Linguistics

• **What Corpus Linguistics Does Not:**
  – Does not explain why. The study of corpora tells us what and how happened but it does not tell us why the frequency of a particular word has increased over time for instance.
  – Does not represent the entire language. Corpus linguistics studies the language by using randomly or systematically selected corpora. They typically consist of a large number of naturally occurring texts, however, they do not represent the entire language. Linguistic analyses that use the methods and tools of corpus linguistics thus do not represent the entire language.
Corpus Linguistics

• Application of Corpus Linguistics:
  – Lexicography. Corpus linguistics plays an important role in compiling, writing and revising dictionaries as within a few seconds, the linguist can get examples of words or phrases from millions of spoken and written texts.
  – Grammar. The huge amount of texts offers a reliable representation of the language to be able to conduct grammatical research as well as to test theoretical hypotheses.
  – Sociolinguistics. Corpus Linguistics offers a valuable insight into how language varies from place to place and between different social groups.
  – Translation studies. Corpora which contain texts in several different languages are a valuable tool for translators as they make it easy to determine how particular words and their synonyms collocate and differ in practical use.
Corpus Linguistics

– **Language learning/teaching.** A growing number of textbooks which are used for language learning/teaching contain texts from corpora rather than “invented” situations because they expose the students to real life situations.

– **Stylistics.** For genres such as the language used by politicians, pop culture, advertising industry, etc., corpora as an important source of information.

– **Dialectology.** The texts included in corpora are in their original form, including dialect which gives the linguists a priceless insight into geographical variation of a language.

– **Historical linguistics.** Historical corpora offer an easy access to virtually all known historic books and manuscripts in electronic form.
Corpus Linguistics

• Notable Corpora:
  – Brown Corpus (the Brown Standard Corpus of Present-Day American English). It contains about 500 English texts that total about 1 million words compiled in the 1960s. It is rather small, but it is the first modern and electronically readable corpus.
  – British National Corpus. It consists of a wide range of written and spoken texts in English language, totalling 100 million words. Since 1994.
  – Oxford English Corpus. It is a huge corpus of English language totalling over 2 billion words. The texts included in the corpus are taken from all sorts of sources, ranging from literary works to the language in forums and chatrooms.
Corpus Linguistics

- **American National Corpus.** It is the American English equivalent to the British National Corpus, however, it only contains about 22 million words of American English spoken and written texts, but it is richly annotated. It is being developed since 1990.

- **International Corpus of English.** It consists of a set of corpora which contain variations of English language from countries where English is either the first or official second language. Each set of the International Corpus of English contains 1 million word texts that have been created after the year 1989.

- **Scottish Corpus of Texts and Speech.** The collection of written and spoken texts in Scottish English and Scots after 1940 is available online for free since 2004. In 2007, the corpus reached a total of 4 million words.

- **WaCky** 2 billion words
**Corpus Linguistics**

- Out of the many possible applications of Corpus Linguistics, we will choose lexical semantics (Generative Lexicon, Pustejovsky 1995) and Distributional Semantics (Baroni 2010).
- The course will focus on Formal vs. Distributional Semantics
Reference/Sense distinction

• Frege: Linguistic signs have a reference and a sense:
  – (i) “Mark Twain is Mark Twain” vs. (ii) “Mark Twain is Samuel Clemens”.
  – (i) same sense and same reference vs. (ii) different sense and same reference.

• Both the sense and reference of a sentence are built compositionally.

• Formal Semantics studies “meaning” as “reference”.

• Distributional semantics focuses on “meaning” as “sense” leading to the “language as use” view.
Formal vs. Distributional Semantics

Focus of FS:
Grammatical words:
- prepositions,
- articles, quantifiers,
- coordination,
- auxiliary verbs,
- Pronouns,
- negation

Focus of DS:
Content words:
- nouns,
- adjectives,
- verbs.
Formal Semantics

• Formal semantics gives an elaborate and elegant account of the productive and systematic nature of language.
• The formal account of compositionality relies on:
  – words (the minimal parts of language, with an assigned meaning)
  – syntax (the theory which explains how to make complex expressions out of words)
  – semantics (the theory which explains how meanings are combined in the process of particular syntactic compositions).
Formal Semantics

**Theory of Meaning**

A *theory of meaning* is understood as providing a detailed specification of the knowledge that a native speaker has about his/her own language. [Dummett, 91]

In doing this, a theory of meaning has to provide a way to assign meaning to all the different words in the language and then a mechanism by means of which all these meanings can be combined into larger expressions to form the meaning of phrases, sentences, discourses, and so on.
Formal Semantics

Truth-conditional semantics program

To state the meaning of a sentence we should state which conditions must be true in the world for this sentence to be true.

e.g. Every man loves a woman.

Truth-conditions:

For each member “x” of the set of men, there should be at least one member “y” of the set of women, in such a way that the pair <x,y> is in the relation loves.

Logic:

\( \forall x. (\text{man}(x) \rightarrow \exists y. (\text{woman}(y) \& \text{loves}(x,y))) \)
Frege’s Compositional Semantics

The meaning of the sentence is determined by the meaning of the words of which it is composed, and the way in which these are put together.

The linear order of the words in a sentence hide the role that different kinds of words play in the building of the meaning of the whole.
Formal Semantics

**Syntactic structure**

**Left diagram:**
- S
- VP
- NP
- PN
  - John
  - likes
  - Mary
- like(john, mary)

**Right diagram:**
- S
- VP
- NP
- Det
- Noun
  - Every
  - man
  - likes
  - Mary
- \(\forall x. (\text{man}(x) \rightarrow \text{likes}(x, \text{mary}))\)
Formal Semantics

• Semantic Structure
• Formal Semantics uses Lambda Calculus as a means of combining meaning guided by the syntactic operations.
Formal Semantics

\[ \forall x. (\text{man}(x) \rightarrow \text{likes}(x, \text{mary})) \]

\[ \lambda P. \lambda Q. \forall x. (P(x) \rightarrow Q(x)) \]

\[ \lambda w. \text{man}(w) \]

\[ \lambda x. \lambda y. \text{likes}(x, y) \]

\[ \lambda P. P(\text{mary}) \]
Distribuotional Semantics

• You shall know a word by the company it keeps (Firth);
• The meaning of a word is defined by the way it is used (Wittgenstein).
• This leads to the distribuotional hypothesis about word meaning:
  – the context surrounding a given word provides information about its meaning;
  – words are similar if they share similar linguistic contexts;
  – semantic similarity = distribuotional similarity.
Distributional Semantics

• Examples of similar words:
  – “astronaut” and “cosmonaut”
  – “car” and “automobile”
  – “banana” and “apple” (these two are less similar)
  – “huge” and “large”,
  – “eat” and “devour”

• Not similar:
  – “car” and “flower”,
  – “car” and “pope”
Distributional Semantics

• For example, if one word describes a given situation
  – “I’m on the highway”
• then it is very likely that the other word also describes this situation
  – “I’m in a car”
• Distributional semantics is an approach to semantics that is based on the contexts of words and linguistic expressions in large corpora.
Distributional Semantics

• Take a word and its contexts. By looking at a word's context, one can infer its meaning.

• tasty *tnassiorc*
• greasy *tnassiorc*
• *tnassiorc* with butter
• *tnassiorc* for breakfast
Distributional Semantics

• He filled the *wampimuk*, passed it around and we all drank some ➔ **DRINK**

• We found a little, hairy *wampimuk* sleeping behind the tree ➔ **ANIMAL**
DS accounts for different uses of words (like in Generative Lexicon). Take “brown” for example. Each adjective acts on nouns in a different way:

“In order for a cow to be brown most of its body's surface should be brown, though not its udders, eyes, or internal organs. A brown crystal, on the other hand, needs to be brown both inside and outside. A book is brown if its cover, but not necessarily its inner pages, are mostly brown, while a newspaper is brown only if all its pages are brown. For a potato to be brown it needs to be brown only outside. . . Furthermore, in order for a cow or a bird to be brown the brown color should be the animal's natural color, since it is regarded as being `really' brown even if it is painted white all over. A table, on the other hand, is brown even if it is only painted brown and its `natural' color underneath the paint is, say, yellow. But while a table or a bird are not brown if covered with brown sugar, a cookie is. In short, what is to be brown is dieffrent for dieffent types of objects. To be sure, brown objects do have something in common: a salient part that is wholly brownish. But this hardly suffices for an object to count as brown. A signficant component of the applicability condition of the predicate `brown' varies from one linguistic context to another.” (Lahav 1993:76)
Distributional Semantics

• What happens with brown is replicated by the large majority of adjective-noun combinations. Treating them all like `idioms' would mean to turn the exception into the rule.

• As it is easy to see, many of the problems come from the lexicon of content words, such as nouns, verbs and adjectives, and not from grammatical terms.
Distributional Semantics

• Of course, there have been important attempts to tackle the lexicon problem from the point of view of formal semantics, like Pustejovský's (1995) theory of the Generative Lexicon.

• More recently, Asher (2011) has approached lexical semantics with a theory of predication that uses a sophisticated system of semantic types, plus a mechanism of type coercion.
Distributional Semantics

• However, the problem of lexical semantics is primarily a problem of size: even considering the many subregularities found in the content lexicon, a hand-by-hand analysis is simply not feasible.

• The problem of assigning reasonable (if not exhaustive) syntactic structure to arbitrary, real-life sentences is perhaps equally hard. Here, however, technology has been an important part of the answer: Natural language parsers, that automatically assign a syntactic structure to sentences, have made great advances in recent years by exploiting probabilistic information about parts of speech (POS tags) and syntactic attachment preferences.
Distributional Semantics

• Tasks where DS has been successful:
  – Word similarity,
  – Information retrieval,
  – Question Answering,
  – Entailment, etc.
Distributional Semantics

- Two words are neighbors if they cooccur.
- The cooccurrence count of words w1 and w2 in corpus G is the number of times that w1 and w2 occur:
  - in a linguistic relationship with each other (e.g., w1 is a modifier of w2) or
  - in the same sentence or
  - in the same document or
  - within a distance of at most k words (where k is a parameter)
Distributional Semantics

- corpus = English Wikipedia
- cooccurrence defined as occurrence within $k = 10$ words of each other:
  - $\text{cooc.}(\text{rich},\text{silver}) = 186$
  - $\text{cooc.}(\text{poor},\text{silver}) = 34$
  - $\text{cooc.}(\text{rich},\text{disease}) = 17$
  - $\text{cooc.}(\text{poor},\text{disease}) = 162$
  - $\text{cooc.}(\text{rich},\text{society}) = 143$
  - $\text{cooc.}(\text{poor},\text{society}) = 228$
• \( \text{cooc.}(\text{poor}, \text{silver}) = 34 \), \( \text{cooc.}(\text{rich}, \text{silver}) = 186 \),
• \( \text{cooc.}(\text{poor}, \text{disease}) = 162 \), \( \text{cooc.}(\text{rich}, \text{disease}) = 17 \),
• \( \text{cooc.}(\text{poor}, \text{society}) = 228 \), \( \text{cooc.}(\text{rich}, \text{society}) = 143 \)
Distributional Semantics

• The similarity between two words is usually measured with the cosine of the angle between them.
• Small angle: silver and gold are similar.
Distributional Semantics

• Up to now we’ve only used two dimension words: rich and poor.
• Now do this for a very large number of dimension words: hundreds or thousands.
• This is now a very high-dimensional space with a large number of vectors represented in it.
• Note: a word can have a dual role in word space.
  – Each word can, in principle, be a dimension word, an axis of the space.
  – But each word is also a vector in that space.
Distributional Semantics

• We can compute now the nearest neighbors of any word in this in word space.

• Nearest neighbors of “silver”:

  1.000 silver / 0.865 bronze / 0.842 gold / 0.836 medal / 0.826 medals / 0.761 relay / 0.740 medalist / 0.737 coins / 0.724 freestyle / 0.720 metre / 0.716 coin / 0.714 copper / 0.712 golden / 0.706 event / 0.701 won / 0.700 foil / 0.698 Winter / 0.684 Pan / 0.680 vault / 0.675 jump
Distributional Semantics

• Nearest neighbors of “disease”:
  1.000 disease / 0.858 Alzheimer / 0.852 chronic / 0.846 infectious / 0.843 diseases / 0.823 diabetes / 0.814 cardiovascular / 0.810 infection / 0.807 symptoms / 0.805 syndrome / 0.801 kidney / 0.796 liver / 0.788 Parkinson / 0.787 disorders / 0.787 coronary / 0.779 complications / 0.778 cure / 0.778 disorder / 0.778 Crohn / 0.773 bowel
Distributional Semantics

• Cases where simple word space models fail:
  – Antonyms are judged to be similar: “disease” and “cure”
  – Ambiguity: “Cambridge”
  – Homonymy: ”bank”
  – Non-specificity (occurs in a large variety of different contexts and has few/no specific semantic associations): “person”
Distributional Semantics

• The vectors in our space have been words so far.
• But we can also represent other entities like: phrases, sentences, paragraphs, documents, even entire books.
• Compositionality problem: how to obtain the distribution vector of a phrase?
Distributional Semantics - from words to phrases

• Option 1: The distribution of phrases – even sentences – can be obtained from corpora, but...
  – those distributions are very sparse;
  – observing them does not account for productivity in language.

• Option 2: Use vector product of two or more words to compute the phrase distribution, but...
  – Multiplication is commutative in a word-based model:
    • [[The cat chases the mouse]] = [[The mouse chases the cat]].
  – Multiplication is intersective – problem for non-intersective adjectives:
Distributional Semantics

• Adjective types, Partee (1995)
• **Intersective:** carnivorous mammal
  \[ ||\text{carnivorous mammal}|| = ||\text{carnivorous}|| \cap \||\text{mammal}|| \]

• **Subsective:** skilful surgeon
  \[ ||\text{skilful surgeon}|| \subset \||\text{surgeon}|| \]

• **Non-subsective:** former senator
  \[ ||\text{former senator}|| \neq ||\text{former}|| \cap ||\text{senator}|| \\
  ||\text{former senator}|| \notin \||\text{senator}|| \]
Distributional Semantics

• DS Strengths:
  – fully automatic construction;
  – representationally simple: all we need is a corpus and some notion of what counts as a word;
  – language-independent, cognitively plausible.

• DS Weaknesses:
  – no generative model
  – many ad-hoc parameters
  – ambiguous words: their meaning is the average of all senses
  – context words contribute indiscriminately to meaning;
    
    \[\text{[[The cat chases the mouse]] = [[The mouse chases the cat]].}\]