Natural Language Processing

Unit 1: Semantic analysis

January 2020

Semantic analysis

Definition

The semantic analysis of natural language content starts by reading all of the words in content to capture the real meaning of any text. It identifies the text elements and assigns them to their logical and grammatical role. It analyzes context in the surrounding text and it analyzes the text structure to accurately disambiguate the proper meaning of words that have more than one definition.

Semantic analysis

Semantic technology processes the logical structure of sentences to identify the most relevant elements in text and understand the topic discussed. It also understands the relationships between different concepts in the text. For example, it understands that a text is about “politics” and “economics” even if it doesn’t contain the the actual words but related concepts such as “election,” “Democrat,” “speaker of the house,” or “budget,” “tax” or “inflation.”

Why it's important

- Understand the human language from a computer science perspective
- Useful tasks as Named Entity Recognition, Information Extraction, Sentiment analysis among others.
- Main goal of the Natural language understanding
- Advances in linguistic and psychology science
- Industry applications.

Approaches

Linguistics Approaches

Linguistics Approaches

Various theories and approaches to semantic representation can be roughly ranged along two dimensions:

- formal vs. cognitive
- compositional vs. lexical

Formal theories have been strongly advocated since the late 1960s, while cognitive approaches have become popular in the last three decades, driven also by influences from cognitive science and psychology.

Discourse representation theory

Discourse representation theory (DRT) was developed in the early 1980s by Kamp (1981) in order to capture the semantics of discourses or texts, that is, coherent sequences of sentences or utterances, as opposed to isolated sentences or utterances.

Discourse representation theory

The basic idea is that as a discourse or text unfolds the hearer builds up a mental representation (represented by a so-called discourse representation structure, DRS), and that every incoming sentence prompts additions to this representation. It is thus a dynamic approach to natural language semantics (as it is in the similar, independently developed File Change Semantics (Heim 1982, 1983)).

Discourse representation theory

DRT formally requires the following components:

- A formal definition of the representation language consisting of:
  - A recursive definition of the set of all well-formed DRSs
  - A model-theoretic semantics for the members of this set
- A construction procedure specifying how a DRS is to be extended when new information becomes available

Putejovsky's generative lexicon

Another dynamic view of semantics, but focusing on lexical items, is Pustejovsky's (1991a,b, 1995, 2001) Generative Lexicon theory. He states: "our aim is to provide an adequate description of how our language expressions have content, and how this content appears to undergo continuous modification and modulation in new contexts" (Pustejovsky 2001: 52).

Putejovsky's generative lexicon

Pustejovsky posits that within particular contexts, lexical items assume different senses. For example, the adjective good is understood differently in the following four contexts:

- A good umbrella (an umbrella that guards well against rain)
- A good meal (a meal that is delicious or nourishing)
- A good teacher (a teacher who educates well)
- A good movie (a movie that is entertaining or thought provoking)

He develops "the idea of a lexicon in which senses [of words/lexical items, CG/AS] in context can be flexibly derived on the basis of a rich multilevel representation and generative devices".

Putejovsky's generative lexicon

This lexicon is characterized as a computational system, with the multilevel representation involving at least the following four levels (Pustejovsky 1995: 61):

- Argument structure: Specification of number and type of logical arguments and how they are realized syntactically.
- Event structure: Definition of the event type of a lexical item and a phrase. The event type sorts include states, processes, and transitions; sub-event structuring is possible.

Putejovsky's generative lexicon

This lexicon is characterized as a computational system, with the multilevel representation involving at least the following four levels (Pustejovsky 1995: 61):

- Qualia structure: Modes of explanation, comprising qualia (singular: quale) of four kinds: constitutive (what an object is made of), formal (what an object is—that which distinguishes it within a larger domain), telic (what the purpose or function of an object is), and agentive (how the object came into being, factors involved in its origin or coming about).
- Lexical Inheritance Structure: Identification of how a lexical structure is related to other structures in the lexicon and its contribution to the global organization of a lexicon.

Natural semantic metalanguage

Natural semantic metalanguage (NSM) is a decompositional system based on empirically established semantic primes, that is, simple indefinable meanings which appear to be present as word-meanings in all languages (Wierzbicka 1996; Goddard 1998; Goddard and Wierzbicka 2002; Peeters 2006; Goddard 2008).

Natural semantic metalanguage

The NSM system uses a metalanguage, which is essentially a standardized subset of natural language: a small subset of word-meanings (63 in number, see Table 5.1), together with a subset of their associated syntactic (combinatorial) properties. Table 5.1 is presented in its English version, but comparable tables of semantic primes have been drawn up for many languages, including Russian, French, Spanish, Chinese, Japanese, Korean, and Malay. Semantic primes and their grammar are claimed to represent a kind of "intersection" of all languages.

Natural semantic metalanguage

NSM is a cognitive theory. Its adherents argue that simple words of ordinary language provide a better representational medium for ordinary cognition than the more technical metalanguages used by other cognitive theories, such as Jackendoff's (1990, 2002) Conceptual Semantics or Wunderlich's (1996, 1997) Lexical Decomposition Grammar.

Object oriented semantic

Object-oriented semantics is a new field in linguistic semantics. Although it is rather restricted in its semantic application domains so far—mainly applied to the representation of verbal meaning—it promises to become relevant for NLP applications in the future, and due to the large body of research in computer science a wealth of resources is already available for object-oriented systems in general. There is some overlap with Pustejovsky's ideas, as he himself observes: When we combine the qualia structure of a NP with the argument structure of a verb, begin to see a richer notion of compositionality emerging, one that looks very much like object-oriented approaches to programming. (Pustejovsky 1991a: 427)

Object oriented semantics

The basic motive behind deploying the computational object-oriented paradigm to linguistic semantics is, however, its intuitive accessibility. The human cognitive system centers around entities and what they are like, how they are related to each other, what happens to them and what they do, and how they interact with one another. This corresponds to the object-oriented approach, in which the concept of "object" is central, whose characteristics, relations to other entities, behavior, and interactions are modeled in a rigorous way. This correspondence between object-orientation and cognitive organization strongly suggests the application of object-orientation for the representational task at hand.

Object oriented semantics

Schalley (2004a,b) introduces a decompositional representational framework for verbal semantics, the Unified Eventity Representation (UER). It is based on the Unified Modeling Language (UML; cf. Object Management Group 1997–2009), the standard formalism for object-oriented software design and analysis. The UER adopts the graphical nature of the UML as well as the UML's general architecture. UER diagrams are composed of well-defined graphical modeling elements that represent conceptual categories. Conceptual containments, attachments, and relations are thus expressed by graphical entailments, attachments, and relations.

Approaches

Computational Approaches

Logical approach

Focus

Logical approaches to meaning generally address problems in compositionality,on the assumption (the so-called principle of compositionality,attributed to Frege) that the meanings of supralexical expressions are determined by the meanings of their parts and the way in which those parts are combined.

Logical approach

One such approach uses the so-called "logical form,"which is a representation of meaning based on the familiar predicate and lambda calculi. The semantics, or meaning, of an expression in natural language can be abstractly represented as a logical form. Once an expression has been fully parsed and its syntactic ambiguities resolved, its meaning should be uniquely represented in logical form.

Logical approaches

Text is processed uing predicates logic and the propositions are interpreted as true or false. Logical semantics try to infer the truth of new propositions:

- Contains predicates, quantifiers and vairables
- $Philosopher(a) \Rightarrow Scholar(a)$
- $\forall x, King(x) \wedge Greedy(x) \Rightarrow Evil(x)$
- Polynomial-time inference procedure exists when KB is expressed as Horn clauses: $P_1 \wedge P_2 \wedge P_3 ... \wedge P_k \Rightarrow Q$

Knowledge base enginnering

- 1. Identify the task.
- 2. Assemble the relevant knowledge.
- 3. Decide on a vocabulary of predicates, functions, and constants.
- 4. Encode general knowledge about the domain.
- 5. Encode a description of the specific problem instance.
- 6. Pose queries to the inference procedure and get answers.
- 7. Debug the knowledge base.

Knowledge engineering

- 1. All professors are people.
- 2. Deans are professors.
- 3. All professors consider the dean a friend or don't know him.
- 4. Everyone is a friend of someone.
- 5. People only criticize people that are not their friends.
- 6. Lucy is a professor
- 7. John is the dean.
- 8. Lucy criticized John.
- 9. Is John a friend of Lucy's?

Term Frequency (TF)

Term frequency: Indicates the frequency of the term in a document. It reflects the relative importance of a term in a document, more frequency indicates more presence of the term and therefore more importance:

$$TF(d,t) = \frac{f(d,t)}{max\{f(d,x) : x \in d\}}$$

Inverse Document Frequency (IDF)

Inverse document frequency: Indicates the frequency of the term among all documents in acorpus D. It reflects how common the word is. If the word is present in many documents is a common word and then not too relevant to specify an specific document.

$$IDF(t) = log(\frac{|D|}{1+|\{d \in D : t \in d\}|})$$

TF-IDF

TFIDF mixes the two techniques in order to see how a term is relevant to each document and how is relevant among a whole corpus.

$$TFIDF(d,t) = TF(d,t).IDF(t)$$

Use in document categorization

TFIDF can be used to perform document categorization, if we transfor the documents into a vector representation using the TFIDF scores of the terms, a new document can be modeled as a vector and then apply classifcation and clustering techniques.

Question answering

Mixture of deep and shallow approaches. The deep approaches exploit NPL parsing and relation extraction methods.

- AquaLog: it's a QA system thattakes queries expressed in natural language and an ontology as input and returns answers drawn from one or more knowledge bases (KBs), which instantiate the input ontology with domain-specific information. More: http://technologies.kmi.open.ac.uk/aqualog/

Question answering

Mixture of deep and shallow approaches. The deep approaches exploit NPL parsing and relation extraction methods.

- IBM Watson: is QA system developed in IBM's DeepQA project; it won the American TV quiz show, Jeopardy. "The DeepQA hypothesis is that by complementing classic knowledge-based approaches with recent advances in NLP, Information Retrieval, and Machine Learning to interpret and reason over huge volumes of widely accessible naturally encoded knowledge (or "unstructured knowledge") we can build effective and adaptable open-domain QA systems." More: http://www.research.ibm.com/deepqa/deepqa.shtml

Information extraction

Information extraction

Information extraction (IE), turns the unstructured information embedded in texts into structured data, for example for populating a relational database to enable further processing.

Named entity recognition

Named entity recognition

The task of named entity recognition (NER) is to find each mention of a named entity in the text and label its type. What constitutes a named entity type is task specific; people, places, and organizations are common, but gene or protein names or financial asset classes might be relevant for some tasks.

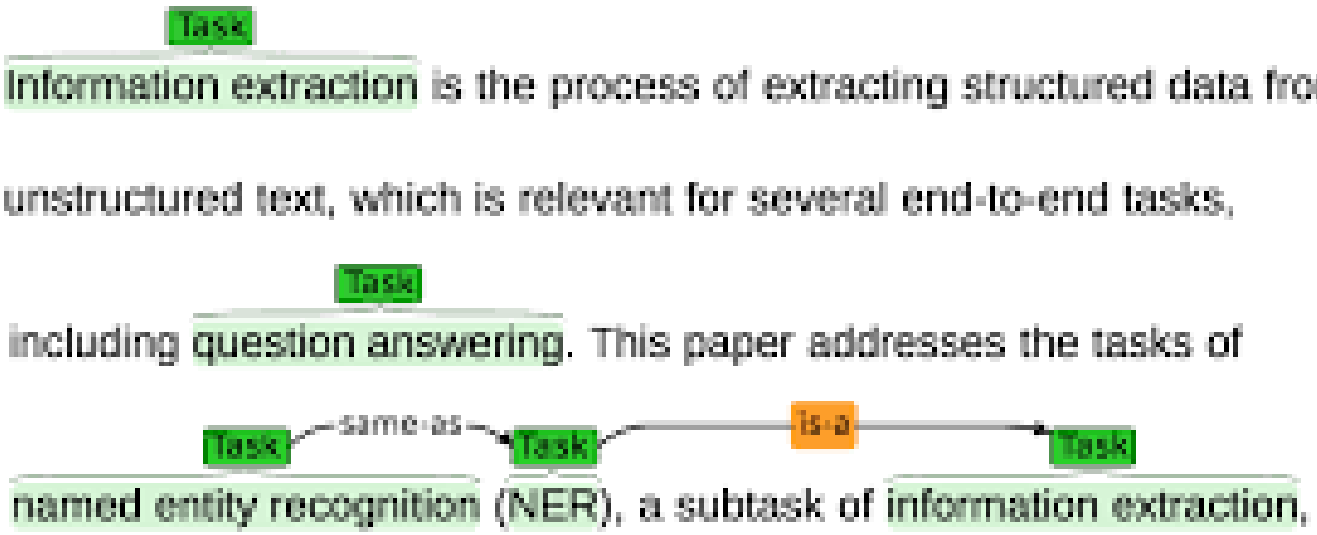
Named entity

A named entity is, roughly speaking, anything that can be referred to with a proper name: a person, a location, an organization. The term is commonly extended to include things that aren't entities per se, including dates, times, and other kinds of temporal expressions, and even numerical expressions like prices. Here's the sample text introduced earlier with the named entities marked

Named entity example

Citing high fuel prices, [ORG **United Airlines**] said [TIME **Friday**] it has increased fares by [MONEY **$6**] per round trip on flights to some cities also served by lower-cost carriers. [ORG **American Airlines**], a unit of [ORG **AMR Corp.**], immediately matched the move, spokesman [PER **Tim Wagner**] said. [ORG **United**], a unit of [ORG **UAL Corp.**], said the increase took effect [TIME **Thursday**] and applies to most routes where it competes against discount carriers, such as [LOC **Chicago**] to [LOC **Dallas**] and [LOC **Denver**] to [LOC **San Francisco**].

Named entity example



Named entity example

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Common categorical ambiguities associated with various proper names.

[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.
The [$_{\text{VEH}}$ Washington] had proved to be a leaky ship, every passage I made...

Relation extraction

Relation extraction

relation extraction: finding and classifying semantic relations among the text entities. These are often binary relations like child-of, employment, part-whole, and geospatial relations. Relation extraction has close links to populating a relational database.

Let's code

Assignments

Assignment 5: Implement a simple LDA analysis of the kaggle toxic comment competition

- Implement a class which apply LDA analysis over the kaggle toxic comments database: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
- The class must perform a pre processing using the methods developed before
- The class must perform the LDA method seen in class.
- Include a jupyter notebook explaining the functionality of the class,

References I

[1] Jurafsky D., Martin J.: Speech and Language Processing 2nd. ed. (2009).

[2] Bird S., Klein E., Loper, E.: Natural Language Processing with Python, (2009). ISBN: 978-0-596-51649-9

[3] Indurkhya N., Damerau F. Handbook of Natural Language Processing, Second Edition (2010). ISBN: 978-1-4200-8593-8

[4] Kao, A., Poteet S. Natural Language Processing and Text Mining (2007). ISBN: 78-1-84628-175-4

[5] Sipser, M. Introduction to the theory of computation (2013). ISBN: 978-1-133-18779-0

[6] https://www.expertsystem.com/natural-language-process-semantic-analysis-definition

References II

[1] *https : //www.tutorialspoint.com/natural_language_processing/natural_language_processing_semantic_analysis.htm*

[2] http://disi.unitn.it/bernardi/RSISE11/Slides/lecture4.pdf

[3] https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[4] https://github.com/kapadias/mediumposts/blob/master/nlp/published_notebooks/Introduction %20to %20Topic %20Modeling.ipynb

[5] https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158

[6] https://medium.com/machine-learning-intuition/document-classification-part-2-text-processing-eaa26d16c719