Natural Language Processing

Introduction

Cristina Tîrnăucă

Dept. Matesco, University of Cantabria

Faculty of Science – Grado en Ing. Informática

Why is it a Hard Task?

English version

- The thieves stole the paintings. They were subsequently sold.
- The thieves stole the paintings. They were subsequently caught.
- The thieves stole the paintings. They were subsequently found.

Spanish translation

- Los ladrones robaron las pinturas. Posteriormente fueron vendidas. Se vendieron posteriormente.
- Los ladrones robaron las pinturas. Posteriormente fueron capturados .
- Los ladrones robaron las pinturas. Posteriormente fueron encontrados.

Some Brief History

- NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence (1956!!!)
- machine translation (MT):
  - XVII-th century - mechanical dictionaries, universal language
  - Weaver and Booth - started one of the earliest MT projects in 1946 on computer translation based on expertise in breaking enemy codes during World War II
  - Weaver, 1949 - "Translation" memorandum, that brought the idea of MT to general notice and inspired many projects
  - early work in MT - simplistic view: the only differences between languages resided in their vocabularies and the permitted word orders.
  - Chomsky, 1957, generative grammars
  - ALPAC report (Automatic Language Processing Advisory Committee of the National Academy of Science - National Research Council) in 1966

Some Brief History, II

- syntactic analysis of phrases in natural language: the depth of ambiguity in the English language
  Time flies like an arrow vs. Fruit flies like a banana.
- Systems prototypes: ELIZA (1964-1966), SHRDLU (1968-1970), PARRY (1972)
- statistical approaches - succeeded in dealing with many generic problems in computational linguistics such as part-of-speech identification, word sense disambiguation, etc.

Major Tasks in Speech Processing

- Speech recognition
- Speech synthesis
- Speaker recognition
- Speech enhancement
- Speech coding
- Voice analysis for medical purposes

Major Tasks in Text Processing

- Machine translation
- Automatic summarization
- Natural language generation
- Natural language understanding
- Question answering
- Information retrieval
- Information extraction and sentiment analysis

Components

- Phonological analysis: phonemes
  This level deals with the interpretation of speech sounds within and across words.
- Morphologic analysis: morphemes (the smallest units of meaning)
  Its goal is to detect the relationship established between the smallest units that make up a word, such as the recognition of suffixes or prefixes. This level of analysis has a close relationship with the lexicon. Usually, the lexicon contains only the root of a word, and it is the morphological analyzer that is responsible for determining whether its gender, number or case are appropriate.
- Lexical analysis: (POS: part-of-speech tagging): noun, adjective, article, pronoun, verb...
  The lexicon is the set of information on every word that the system uses for processing. The words in the dictionary are represented by a lexical entry, and if it has more than one meaning or different grammatical categories, will be assigned different entries.
  Morphological information, grammatical category, syntactic irregularity and meaning representation is included in the lexicon. Also, at this level we can replace those words that only have one meaning with their semantic representation.
- Syntactical analysis (grammar, parser): subject, predicate, direct and indirect objects,...

Components, II

- This level is responsible for labeling each of the syntactic components that appear in a phrase and analyzing how words combine to form correct grammatical structures. The result of this process consists in generating a derivation tree
- Semantical analysis
  One must distinguish between context free and context sensitive meaning. The context free meaning of a word (the one treated by semantics), refers to the meaning that words have by their own, ignoring the influence of the context or the speaker's intentions. Context sensitive meaning (studied by pragmatics) refers to the meaning of words in certain circumstances
- Discourse analysis: anaphora, cataphora...
  While previous components work at phrase level, discourse analysis must take into consideration a much more complex context (it studies how previous knowledge is relevant to the text under analysis)
- Pragmatic analysis: insinuations, allusions, wordplay, presuppositions...
  Pragmatic analysis adds additional information to the analysis of the meaning of the phrase in a given context.

Moreover, one can include other levels of knowledge such as knowledge of the world, referring to the general knowledge people must have about the structure of the world in order to maintain a conversation.

Difficulties in NLP

- Ambiguity (at lexical level, referential level: anaphora and cataphora, structural level: grouping or functional, pragmatic level,...)
  Yesterday, I found a bow in the garden. 
  The thieves stole the paintings. They were subsequently found.
  Because he was very cold, David put on his coat.
  "One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know." - Groucho Marx
  Visiting relatives can be boring.
  "Salió de la cárcel con tanta honra, que le acompañaron doscientos cardenales sino que a ninguno llamaban eminencia." - Francisco de Quevedo
- Disambiguation: word-category disambiguation (POS), word-sense disambiguation (WSD)

Difficulties in NLP, II

- Text segmentation: word segmentation, sentence segmentation
  In spoken language it is unusual to make a pause between words. Very often, identifying the boundaries between two words must take into account context, grammar and semantics. In written language, languages like Mandarin do not have spaces between the words.
- Imperfect data reception
  Foreign accents, regionalisms or difficulties in pronunciation, typing errors or ungrammatical expressions, errors in reading texts by OCR, etc.

Grammars for NLP, a retrospective

- The structuralist approach (Ferdinand de Saussure, 1920-1950) in Europa or constituency approach (Leonard Bloomfield) in US.
  natural language = a structure of mutually linked elements, similar or opposed to each other
- Chomsky: generative grammars (CFGs,...)
- Transformational grammars (Chomsky)
- Valencies
- Constraints: generalized phrase structure grammars (GPSG)
- Head-driven phrase structure grammars (HDPSG)
- Unification
- Meaning-text theory (MTT) - dependency grammars

NLP today

- "Trade-off" between grammars with nice properties and grammars that are easy to parse from a computational point of view
- Modern parsers are, at least partially, statistical: they are based on corpora with labeled training data
  Colorless green ideas sleep furiously. / Furiously sleep ideas green colorless.
- The vast majority of modern statistical algorithms use some (modified) form of chart parsing
  - Examples: Earley parser, Cocke-Younger-Kasami (CYK) parser
  - suitable for ambiguous grammars
  - dynamic programming
- Tools for machine translation: tree transducers, synchronous grammars, tree bimorphisms