Speech and Language Processing

An introduction to the Natural Language Processing course

Ing. R. Tedesco. PhD, AA 19-20

(mostly from: Speech and Language Processing - Jurafsky and Martin)

Why Should You Care?

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Major Topics

1. Words
2. Syntax
3. Meaning
4. Discourse
5. Speech

6. Applications exploiting each

Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse processing
- Dialogue structure

Topics: Techniques

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Unification
  - Lambda calculus
- First order logic

- Probability models
- Supervised machine learning methods
- Neural Networks

Topics: Applications

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Stand-alone
- Enabling applications
- Funding/Business plans

Categories of Knowledge

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse
- Prosody

Each kind of knowledge has associated with it an encapsulated set of processes that make use of it.

Interfaces are defined that allow the various levels to communicate.

This usually leads to a pipeline architecture.

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

Ambiguity

- Find at least 5 meanings of this sentence:
  - I made her duck
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
  - **Lexical category**: "duck" can be a N or V
- I cooked waterfowl belonging to her
  - **Lexical category:** "her" can be a possessive ("of her") or dative ("for her") pronoun
- I made the (plaster) duck statue she owns
  - **Lexical Semantics:** "make" can mean "create" or "cook"

Ambiguity is Pervasive

- **Grammar**: "Make" can be:
  - **Transitive: (verb has a noun direct object)**
    - I cooked [waterfowl belonging to her]
  - **Ditransitive: (verb has 2 noun objects)**
    - I made [her] (into) [undifferentiated waterfowl]
  - **Action-transitive (verb has a direct object and another verb)**
  - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
  - I mate or duck
  - I'm eight or duck
  - Eye maid; her duck
  - Aye mate, her duck
  - I maid her duck
  - I'm aid her duck
  - I mate her duck
  - I'm ate her duck
  - I'm ate or duck
  - I mate or duck

Brief history - 1

Starting research fields:

- Linguistics
- Natural Language Processing (computer science)
- Speech Recognition (electronics)
- Computational Linguistics (psychology)

1940-1950 - World War II

- Finite State Automata: Formal Language Theory (algebra and set theory for the formalization of languages) - Chomsky (56), Backus (59) and Naurs (60)
- Probabilistic algorithms for speech, information theory (Shannon), noise of the channel encoding and decoding, entropy of a language
- Machine Translation is the most desired application

Brief history - 2

1957-1970 - Two paradigms

- **Symbolic**
  - a) Formal Language Theory (Chomsky): parsing algorithms (first top-down and bottom-up then with dynamic programming)
  - b) Artificial Intelligence - Logic Theories - (from Newell and Simon) - a combination of pattern matching and search for keywords with simple heuristics for reasoning and answer questions
  - a) and b) lead to the early systems
- **Stochastic**: Bayesian method and use of dictionaries and corpora (the first OCR) - Browning (59), Mosteller and Wallace (64). The Brown Corpus (63-64) - Kucera and Francis (67)

Brief history - 3

1970-1983 – FS Models

- Understanding natural language - Winograd 72 (the SHRDLU parser and construction of a systemic grammar): parsing well understood
- You could start working seriously on semantics and discourse (Schank et al 77: scripts, plans and goals, human memory) (Quillian (68), Rumelhart and Norman (75), Simmons 73, ...) with network-based semantics integrated 'case roles' (Fillmore 68)
- Discourse Modeling
  - Analysis of substructures of discourse (Grosz 77, Sidner 83)
  - Automatic resolution of references (Hobbs78)
  - Belief-Desire-Intention (Perrault, Allen 80 - Cohen and Perrault 79)

Brief history - 4

1983-1993 - empiricism and FS Models

- Continuation of finite-state models
  - For phonology and morphology (Kaplan and Kay - 81)
  - For syntax (Church - 80)
- Return to empiricism
  - Work at IBM for speech recognition based on probabilistic models
  - Data-driven approaches: POS tagging, parsing and annotation, for ambiguity resolution, use of connectionist models... from speech recognition to the semantics
- Natural Language Generation

Brief history - 5

1994-1999 - decline of symbolic approach

- Difficulties with symbolic approach to improve
- Heavy use of data-driven methods and probabilistic models
- Enlargement of the application fields (from the Web to Alternative and Augmentative Communication...)

2000-2010 - empiricism and Machine Learning

- The empirical approach becomes even more significant
  - A lot of material written and talked about a lot of material available online and already 'annotated' (in terms of syntactic, semantic and pragmatic aspects)
- Close liaison with the research community of 'machine learning'
  - Focus on learning
  - New opportunities relied to high-performance computing resources
  - Unsupervised systems become more important than the first favorite supervised systems (the trend is set to grow further)

Brief history - 6

2010-2018 - Machine Learning everywhere

- Neural Networks for NLP
  - NN-based ASR/TTS, …
- Conversational Agents
- Emotion and Affect
- Subjectivity and Sentiment Analysis
- Personality

General info on the NLP course

- Lectures:
  - Monday 16:30 – 18:00
  - Wednesday 10:30 – 13:00 (with 15 min break)
  - Thursday 10:30 – 13:00 (with 15 min break)

- Four hands-on lab sessions will give you extra points; signatures will be collected. NB: Dates could change...
  - Thursday 2019-9-26
  - Monday 2019-9-30
  - Wednesday 2019-10-9
  - Monday 2019-10-28

- Exam:
  - Written: 3 topics, with 3 open questions each → max mark: 30
    - At least 18 is required!
    - Could require to solve simple numeric exercises; no calculator is needed
  - Labs: max 2 points
  - Final mark: written part mark + lab points

General info on the NLP course

- Tools:
  - NLTK: `https://www.nltk.org`
  - Praat: `http://www.fon.hum.uva.nl/praat`
- Web site: `http://corsi.dei.polimi.it/nlp`
- **Usually**, slide posted **before** the lecture