

IN THE NAME OF GOD

Natural Language Processing

منابع

- Christopher D. Manning and Hinrich Schutze, Foundation of Statistical Natural Language Processing, MIT Press, 2001.
- Papers
- Daniel Jurafsky and James H. Martin, Speech and language processing: An Introduction to Natural Language Processing, computational Linguistic, and speech Recognition, Second Eddition, Prentice Hall, 2009.
- Others:
- Steven Bird, Ewan Kein, and Edward Loper, Natural Language Processing with Python-Analysing Text with Natural Language Toolkit, OReily Media, 2009.
- James Allen, Natural Language Understanding, Adison Wesley, 1994.

Journals & conferences

- Journals:
- Computational linguistics, TACL, NL engineering, information retrieval, information processing and management, ACM trans. On IS, ACM TALIP, ACM TSLP
- Conferences:
- ACL/NAACL/EMNLP, SGIK, AAJ/IJCAL, collig HLT, \EACL, NAACL, AMTA/MT summit, ICSLP/Europe speech

NLP applications around us

- Suggest in search
- Automatic gmail replies
- Machine translation
- Conversation systems

NLP tasks

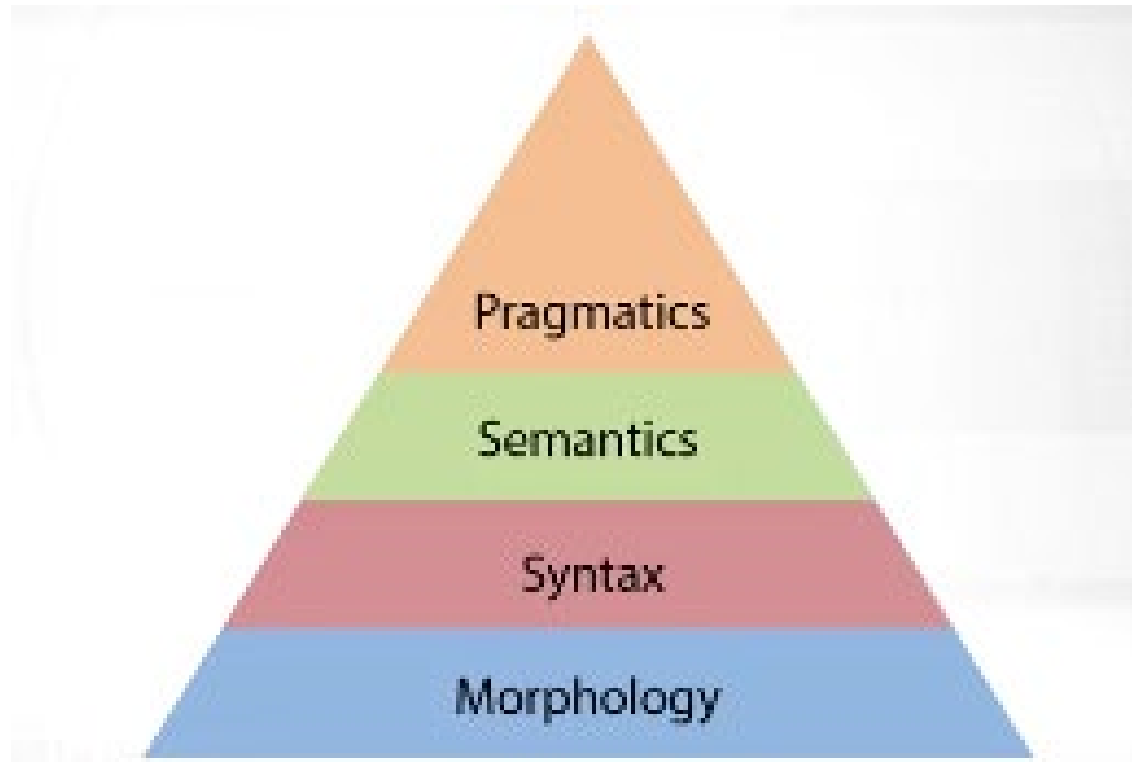
- Text classification
- Named entity recognition
- Duplicates detection
- ...

Problems with text processing

- **Ambiguity**
- Morphological: impossible- important
- Phonetical: singer, finger
- POS: joe wins the first round
- Syntax: Call joe a taxi
- PP statement: joe ate a pizza with a fork/meatballs/ pleasure
- Cc attachment: joe likes ripe apple and pear
- Negation: joe likes pizza with no cheese and tomato
- Referential: Joe yelled at Mike. He has broken the bike. He was angry of him.

- Reflexive: He bought him (himself) a present.
- Paralellism/ellipsis: joe gave Mike a bear and Jeremy a glass.
- **Uncommon words (tel. No,,..)**
- **Unk words(not in dictionary. E.g. Selfie, chillax)**
- **inconsistency (junior college, college junior)**
- **Polysemy (plant, شیر)**
- **Parsing problem مجلس شورای اسلامی**
- ...

NLP pyramid- Levels of text processing



outline

- Words (Text classification tasks:
 - Predict some tags or categories
 - Analyze Sentiment of a review
 - Filter spam e-mails...
- Seq. of words
 - Predict word seq .(language models) used in chat-bots, speech recognition, MT, summarization...
 - Predict tags of word sequences (e.g. POS tags, named entities, semantic slots)

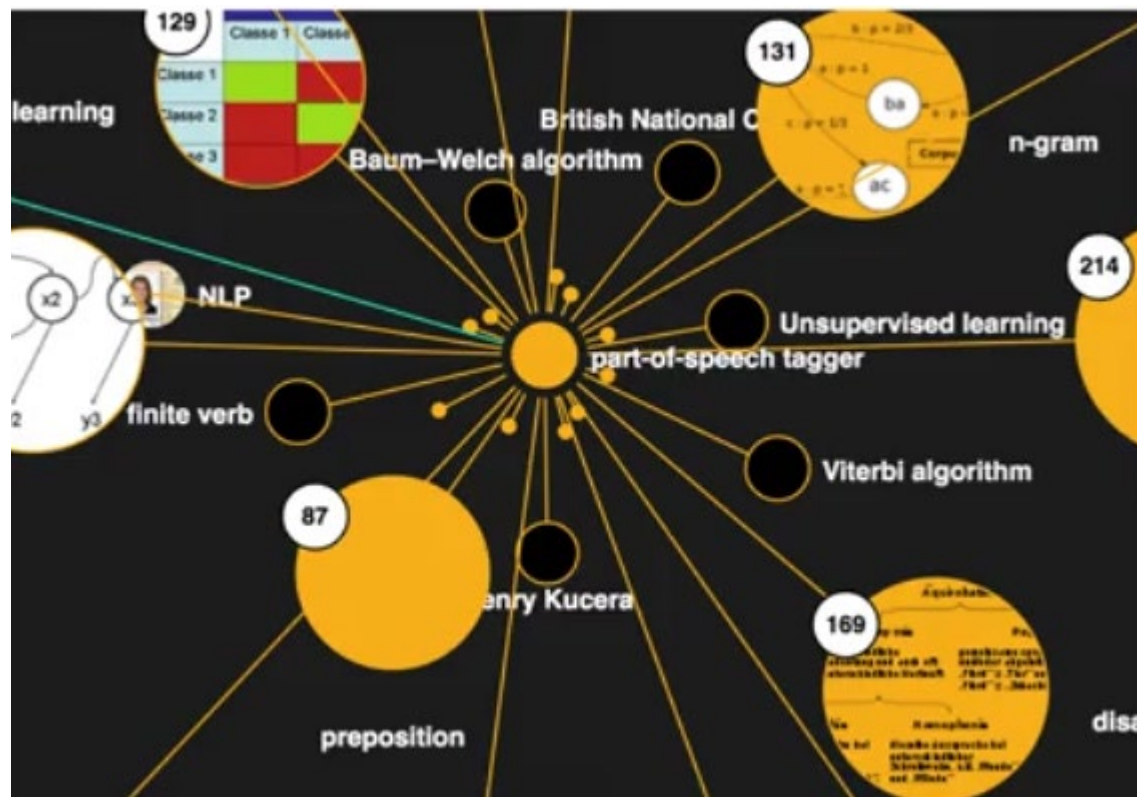
- Meaning of words
 - WordNets
 - Word/sentence embedding
 - Topic models
- Seq to seq
 - MT
 - summarization, simplification
 - Conversational chat-bot

Libraries and tools

- NLTK
 - Preprocessing tools: tokenization, normalization,...
 - Pre-trained models for pos-tagging, parsing,...
- Stanfordparser
- spaCy: lib for NLP
- Gensim: word embedding , topic modelling
- MALLET: java lib for classification, seq tagging, topic modelling

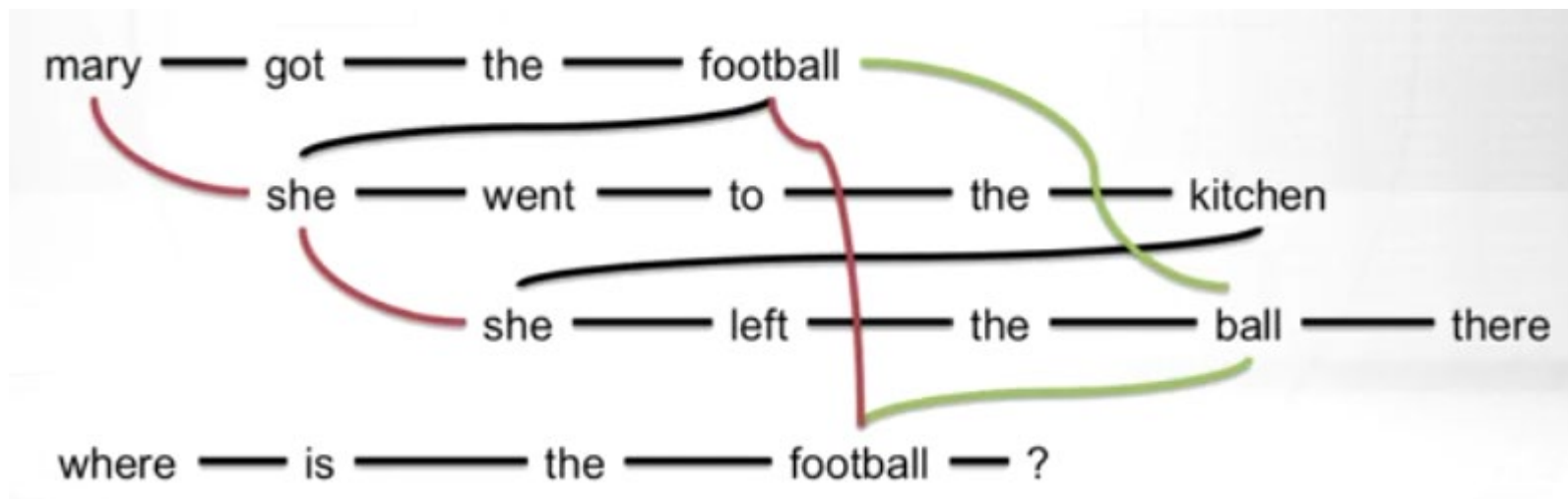
Linguistic knowledge

- Ideas & evaluation
- External resources: wordNet, babelNet, etc.
- <http://babelnet.org/synset?word=NLP&lang=En&details=1&orig=NLP>



Linguistic knowledge+DL

- Coreference resolution, hypernyms (used in QA/reasoning)
- Method: DAG-LSTM
- Linguistic knowledge as memory for recurrent neural networks, 2017

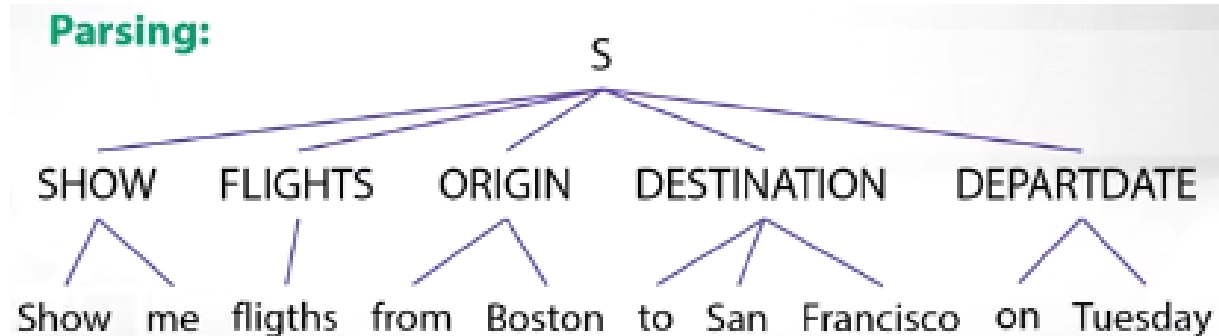


Approaches inside

- Rule based
 - Regular expressions
 - Context-free grammars
 - ..
- Probabilistic modelling and ML
 - Likelihood maximization
 - Linear classifiers
 - ...
- Deep learning
 - Recurrent neural networks...

example-semantic slot filling- 1- rule based

- CFG:
 - S-> Show Flights Origin Destination DepartDate
 - Show-> show me| I want| can I see| ...
 - Flights->(a)flight|flights
 - Origin->from City
 - Destination -> to City
 - City -> Boston| San Fransisco|...
 - DepartDate -> on Day
 - Day -> Saturday,...



example-semantic slot filling-2-probabilistic model (CRF)

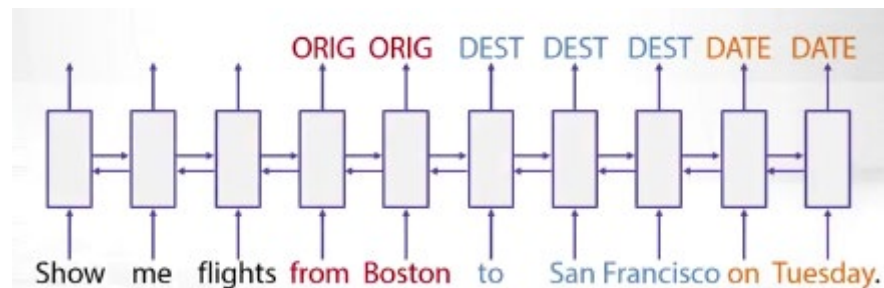
- $P(\text{tags}|\text{word}) = \dots (\text{features}, \text{params } \theta)$
- Train: $p(\text{tags}|\text{words}) \rightarrow \max \theta$
- Inference: $\text{tags}^* = \operatorname{argmax} p(\text{tags}|\text{words})$

example-semantic slot filling-3-ML

- Training corpus:
 - **Orig** **Dest** Date
 - Show me flights **from Boston** **to San Francisco** on Tuesday.
- Feature engineering:
 - Is the word capitalized, is it a city name, previous word, previous slot

example-semantic slot filling-4-DL(LSTM)

- Big training corpus
- No feature generation
- Defining the model
- Training and inference



DL vs. traditional approaches

- Traditional
 - Perform good enough in many tasks (e.g. seq labeling)
 - Allow us not to be blinded with the hype (e.g. word2vec)
 - Can help to further improve DL models (e.g. word alignment prior to MT)
- DL
 - State-of-the-art performance in many tasks (e.g. MT)
 - This is where the most of research in NLP is now happening