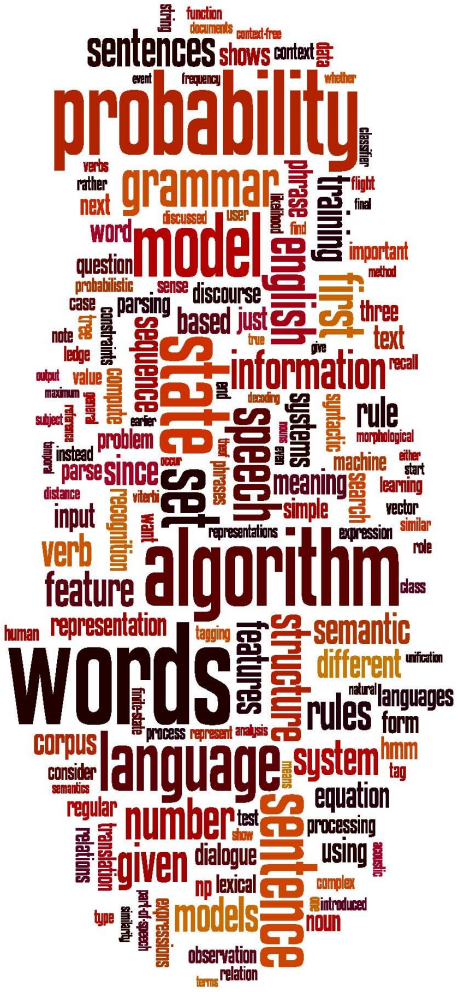
Dan Jurafsky



Natural Language Processing

Lê Anh Cường

lacuong@it.tdt.edu.vn

Content

1. What & Why NLP
2. Why is NLP hard
3. Problems in NLP
4. Approaches in NLP
5. Syllabus

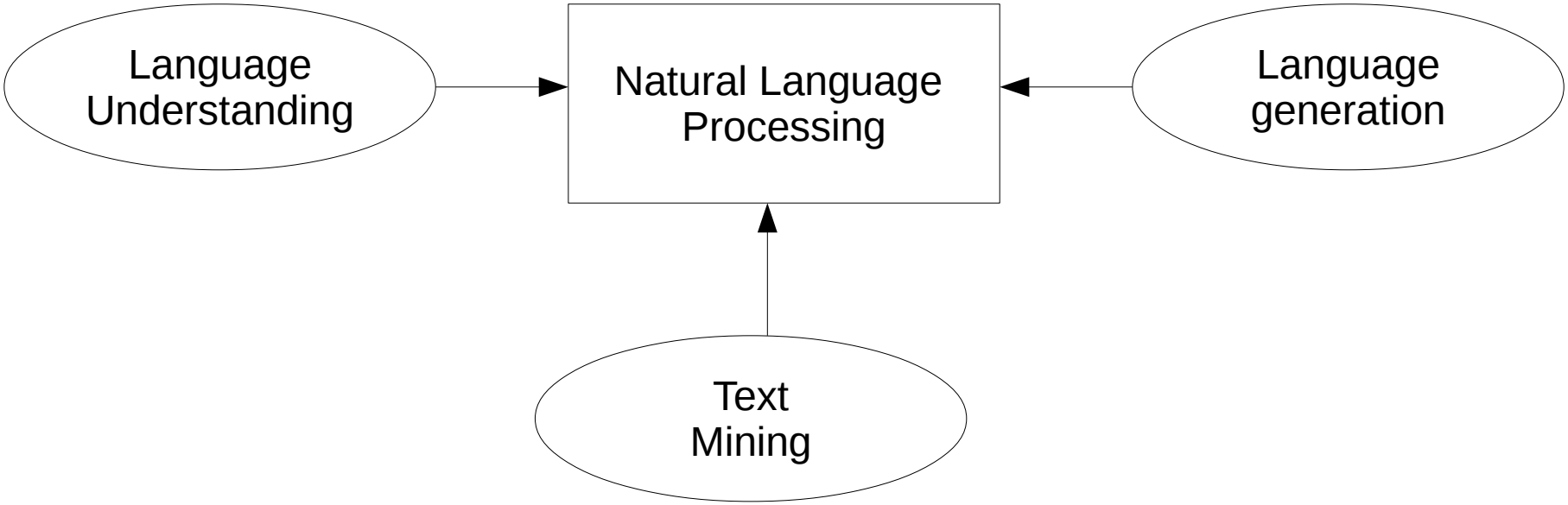
Why NLP?

- Text is everywhere, contains almost information
- Text is the way of communication

Why NLP?

- How does human can communicate with computer?
- How can we get knowledge from texts?

What is NLP?


Language
Understanding
Natural Language
Processing
Language
generation
Text
Mining

Concepts

- Computational Linguistics
- Natural Language Processing

Why is NLP hard?

- Natural langue is ambiguous!

Why is NLP hard?

- Ambiguity

At all levels: lexical, phrase, semantic

Iraqi Head Seeks Arms

Word sense is ambiguous (head, arms)

Stolen Painting Found by Tree

Thematic role is ambiguous: tree is agent or location?

I saw the man with the telescope

At last, a computer that understands you like your mother

Syntactic structure (attachment) is ambiguous: is “the telescope” link to “saw” or “the man”

Hospitals Are Sued by 7 Foot Doctors

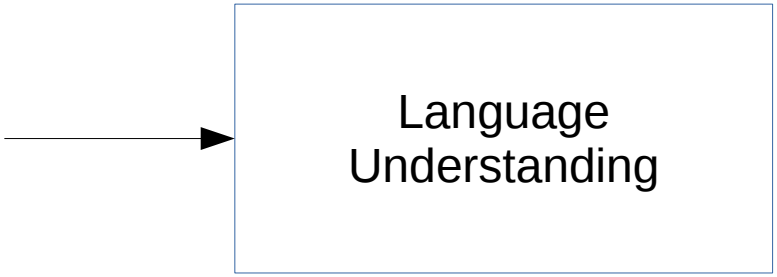
Semantics is ambiguous : what is 7 foot?

Why is NLP hard?

- Rules but exceptions
- Lack of context: World Knowledge

Problems in NLP

- Language model
- Morphology analysis
- Word Segmentation
- Part-Of-Speech tagging
- Syntactic parsing
- Word sense disambiguation
- Semantic representation
- Corereference resolution
- Preposition attachment
- WordNet construction



Applications in NLP

- Spelling

NLP applications

- Spelling
- Grammar checking
- Machine Translation
- Question Answering
- Text catergorization/classification
- Information Retrieval
- Information Extraction
- Opinion Mining & Sentiment Analysis

Approaches in NLP

- Knowledge Based Approach

  For example: Machine Translation

Approaches in NLP

- Statistical approach
  - Empirical methods
  - Data-driven methods

For example: Machine Translation

Approaches in NLP

- Hybrid approach
  - Linguistical knowledge
  - Data
  - Machine learning

For example: Machine Translation

Current approach

- Deep Learning
  - Representation learning

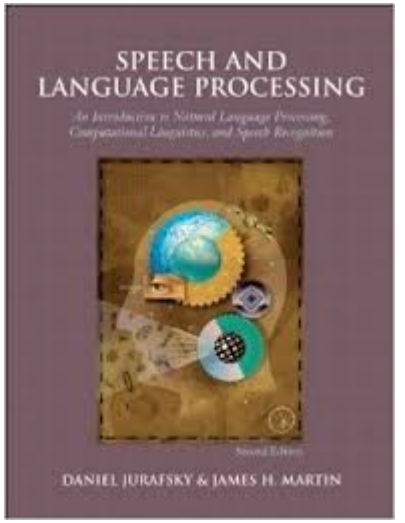
    for example: Word2Vec models

Syllabus

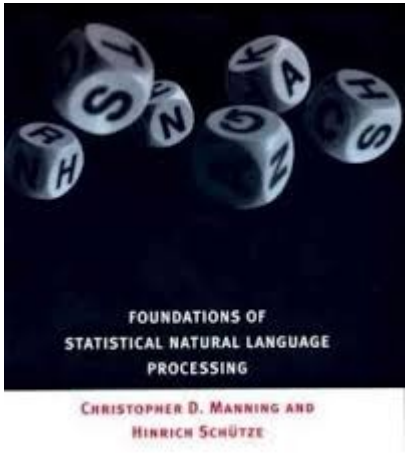
1) Introduction + Language Model
2) Basic Models: Naive Bayesian Classification, Hidden Markov Model
   - Text classification/categorization
   - Part-Of-Speech taggin
3) Maximum Entropy Model + EM algorithm
   * Project 1 report
4) Syntactic parsing
   - CYK algorithm, Earley algorithm, Dependency parsing.
5) Statistical Machine Translation
   *Project 2 report
6) Semantic parsing
   - Word Sense Disambiguation; Semantic Role Labeling; WordNet
7) Introduction to Deep learning models
   - Word Embedding; Recurrent Neural Network
8) Seminar
   - Question Answering; Information Retrieval; Text Summarization; Computational Discourse; Information Extraction
9) Seminar (continue)
10) Final exam: Show experiment

Books

1) Daniel Jurafsky and James H. Martin. 2008. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Second Edition. Prentice Hall.



2) Christopher D. Manning and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press.



Assessment

- 2 assignments
- Middle exam
- Final project

Project 1

- Build a language model and apply to fill in the correct word to the blank place in a text.
- Step 1
- Step 2
- Step 3

Project 2

- Build HMM model and apply to:
  - Vietnames Word Segmentation
  - Part-Of-Speech tagging
  - Name Entity Recognition

Final project

- Study a topic/problem
- Survey related papers
- Presentation
- Do experiments