Introduction to Natural Language Processing

Lecture 1. Introduction

Ekaterina Chernyak, Dmitry Ilvovsky

echernyak@hse.ru, dilvovsky@hse.ru

National Research University
Higher School of Economics (Moscow)

July 14, 2015

Brief history of NLP

- January 7, 1954 – Georgetown experiment. Russian to English machine translation;
- 1957 – Noam Chomsky introduced “universal grammar”;
- since 1961 – Brown Corpus;
- the late 1960’s – ELIZA, a simulation of a psychotherapist;
- 1975 – Vector Space Model by Salton;
- up to the early 1980’s – rule based approaches;
- after the early 1980’s – machine learning, corpus linguistics;
- 1998 – Language Model by Ponte and Croft;
- since 1999 – topic modeling (LSI, pLSI, LDA, etc);
- 1999 – “Foundations of Statistical Natural Language Processing” by Manning and Shuetze;
- 2009 – “Natural Language Processing with Python” by Bird, Klein, and Loper.

Major tasks of NLP

- Machine Translation
- Text classification
  - Sentiment analysis
  - Spam filtering
  - Classification by topic or by genre
- Text clustering
- Named entity recognition
- Question answering
- Automatic summarization
- Natural language generation
- Speech recognition
- Spell checking
- User study design and evaluation

NLP techniques

- The level of characters:
  - Word segmentation
  - Sentence breaking
- The level of words – morphology:
  - Part of speech (POS) tagging
  - Word sense disambiguation
- The level of sentences – syntax:
  - Parsing
- The level of senses – semantics:
  - Coreference resolution
  - Discourse analysis
  - Semantic role labeling
  - Synonymy detection

Main problems

- Ambiguity
  - Lexical ambiguity:
    - Time flies like an arrow; fruit flies like a banana.
  - Syntactic ambiguity
    - Police help dog bite victim.
    - Wanted: a nurse for a baby about twenty years old.
- Neologism: unfriend, retweet, instagram
- Different spelling: NY, New York City, New-York
- Non-standard language: HIIII, how are u? miss u SOOOO much:((((

About this course

We will cover the following topics:

- Tokenization
- POS tagging
- Key word and phrase extraction
- Parsing
- Synonyms detection
- Language sources
- Topic modeling
- Text visualisation

We will try to use Python and R for various tasks.

Further reading