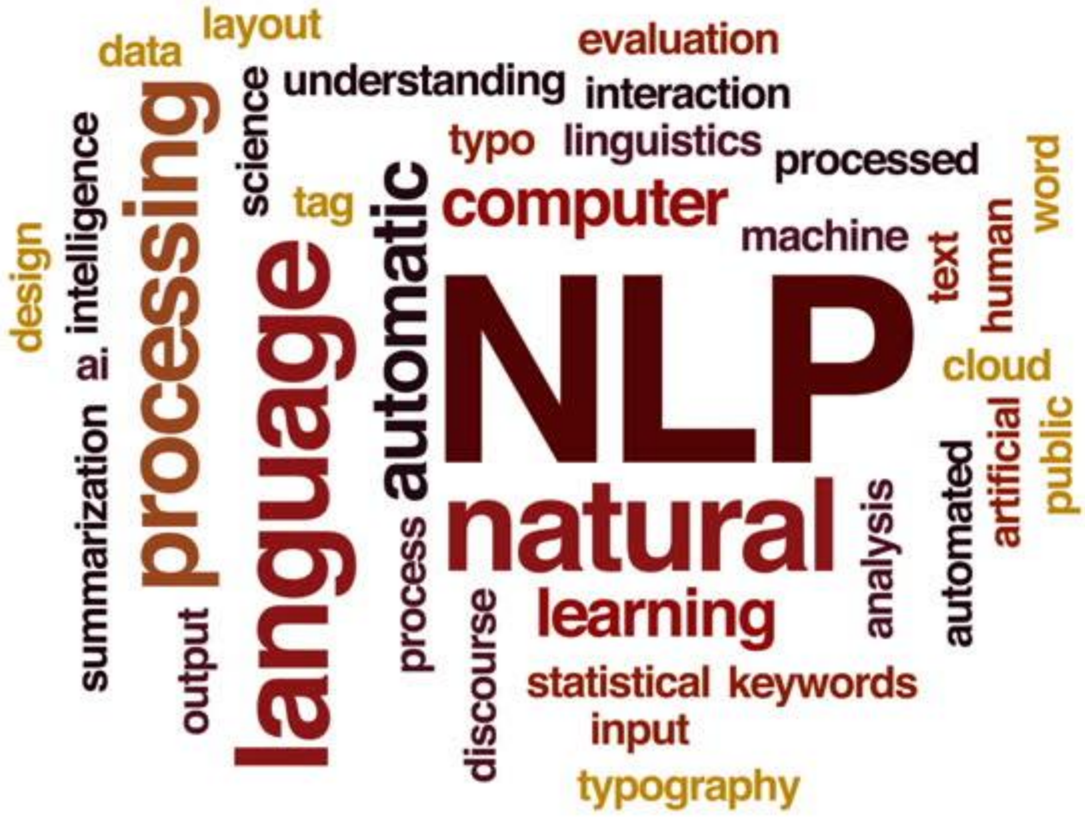
Natural Language Processing

Introduction


NLP
natural
language
processing
automatic
computer
learning
layout
data
evaluation
understanding
interaction
typo
linguistics
processed
science
tag
machine
word
human
text
cloud
public
artificial
automated
analysis
intelligence
design
ai
summarization
output
process
discourse
statistical
keywords
input
typography

What is Natural Language Processing (NLP)?

NLP is a field at the intersection of:

- computer science
- artificial intelligence
- linguistics

Goal: for computers to process or “understand” natural language in order to perform tasks that are useful

Сложности в русском языке

- **Анафоры**

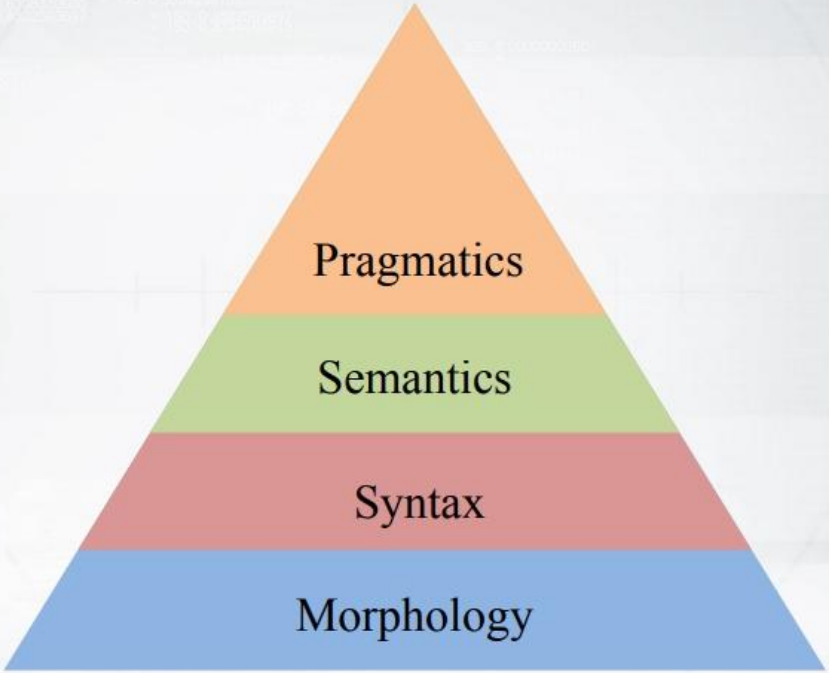
  «Мы отдали бананы обезьянам, потому что они были голодные» и «Мы отдали бананы обезьянам, потому что они были перезрелые»

- **Свободный порядок слов (компенсируется** морфологией, служебными словами и знаками препинания)

  «Бытие определяет сознание» — что определяет что?

- **Неологизмы**

- **Омонимы**



Pragmatics

Semantics

Syntax

Morphology

Natural Language Processing Pyramid

Задачи

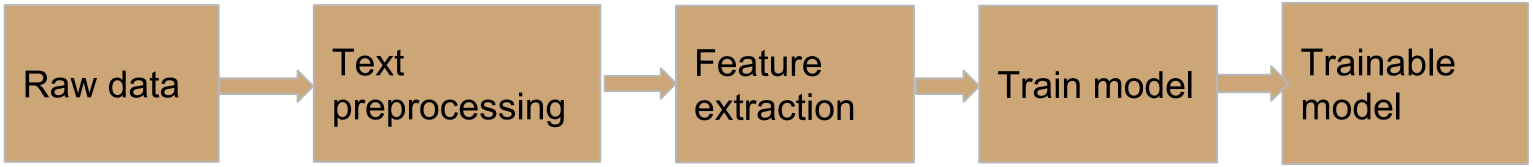
- Machine Translation
- Question Answering
- Dialog System or Conversational Agent (CA)
- Sentiment Analysis
- Speech Recognition
- Text Summarization
- Text Classification
- Optical Character Recognition
- Named Entity Extraction
- Semantic Text Similarity
- Topic modeling

(Stemming, Lemmatization, Part of Speech Tagging, Named Entity Recognition, Coreference resolution, Syntactic Parsing, Word sense disambiguation)

Main approaches in NLP

1. Rule-based methods
   - Regular expressions
   - Context-free grammars
   - ...
2. Probabilistic modeling and machine learning
   - Likelihood maximization
   - Linear classifiers
   - …
3. Deep Learning
   - Recurrent Neural Networks
   - Convolutional Neural Networks
   - …

Pipeline



Raw data → Text preprocessing → Feature extraction → Train model → Trainable model

Text Preprocessing

- Tokenization
- Token Normalization (Stemming, Lemmatization)
- Normalizing Capital Letters, Acronyms
- Noise Removal
- Token Standardization (replacing tokens)

Feature Engineering on text data

- Bag of Words
- N-Grams as Features
- Tf-idf

Bag of Words Example

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

Stopword List
for
is
of
the
to

Problems BOW

- Loose word order
- counters are not normalized

Complex Features

- Semantic features (Word Embeddings, Topic Modeling, NER, ...)
- Syntactic features (Dependency tree, consistency tree, ...)
- Morphological features (PoS, …)

Notebook example

Text classification with 4 classes using BOW as features and 3 different models: logistic regression, random forest classifiers, simple neural network with 2 dense layers.

Metrics - Accuracy