Part-of-speech tagging.[1]

Victor Kitov

v.v.kitov@yandex.ru

[1] With materials used from "Speech and Language Processing", D. Jurafsky and J. H. Martin.

Usage of part of speech

- Parts of speech (also known as word classes, syntactic categories) useful, because:
  - we can predict parts-of-speech of neighbouring words
    - better language model
  - POS is first step in syntactic analysys
    - e.g. noun is part of noun phrase
  - improve named entity recognition
    - names, organizations are usually nouns
  - improve sentiment analysis
    - adjectives describe polarity of comment
  - improve summarization
    - nouns are most descriptive

Categorization of part of speech

- POS may be:
  - open class types (new representatives frequently appear)
    - e.g. nouns: new words iPhone, fax, etc.
    - verbs, adjectives, adverbs also belong to this category
  - closed class types (words in the class are more or less fixed)
    - e.g. function words, such as of, it, and, or you.

Open class types

- noun (существительное)
  - proper nouns (имена собственные)
  - common nouns (нарицательные)
    - count nouns (e.g. goat, relationship)
    - mass nouns (e.g. snow, salt, communism)
- verb (глагол)
  - non-third-person-sg (eat), third-person-sg (eats), progressive (eating), past partici- ple (eaten)
- adjective (прилагательное)
- adverb (наречие)
  - locative adverbs (home, here, downhill)
  - degree adverbs (extremely, very, somewhat)
  - manner adverbs (slowly, slinkily, delicately)
  - temporal adverbs (yesterday, Monday)

Closed class types in English

- prepositions (предлоги): on, under, over, near, by, at, from, to, with
- determiners (артикли): a, an, the
- pronouns (наречия): she, who, I, others
- conjunctions (союзы): and, but, or, as, if, when
- auxiliary verbs (вспомогательные глаголы): can, may, should, are
- particles (частицы): up, down, on, off, in, out, at, by
- numerals (числительные): one, two, three, first, second, third
- + some others

Comments

- Phrasal verb - when verb and particle form together form new meaning unit
  - e.g. turn down=reject, find out=discover, etc.
  - the same in Russian?
- Pronouns (местоимения)
  - Personal pronouns (she, I, it, me)
  - Possessive pronouns (my, your, his, her, its)
  - Wh-pronouns (what, who, whom, whoever)

tag types, corpora

- Different notations may be used for POS.
- Penn treebank is most popular for English
- POS labelled corpora exist in the format like:
  - The/**DT** grand/**JJ** jury/**NN** commented/**VBD** on/**IN** a/**DT** number/**NN** of/**IN** other/**JJ** topics/**NNS** ./**.**

Famous labelled corpora

- **The Brown corpus** is a million words of samples from 500 written texts from different genres published in the United States in 1961.
  - different styles
- **The WSJ corpus** contains a million words published in the Wall Street Journal in WSJ 1989.
  - news, mostly about finance
- **The Switchboard corpus** consists of 2 million words of telephone conversations collected in 1990-1991.
  - conversations

POS tagging

- Part-of-speech tagging - the process of assigning a part-of speech marker to each word in an input text.
- First tokenization should be made
- Tagging should resolve ambiguities:
  - book that flight, book=verb
  - give me that book, book=noun
- Number of ambiguities in Wall Street Journal and Brown corpus[2]:

		WSJ	Brown
Types:			
Unambiguous	(1 tag)	44,432 (**86%**)	45,799 (**85%**)
Ambiguous	(2+ tags)	7,025 (**14%**)	8,050 (**15%**)
Tokens:			
Unambiguous	(1 tag)	577,421 (**45%**)	384,349 (**33%**)
Ambiguous	(2+ tags)	711,780 (**55%**)	786,646 (**67%**)

[2] types are unique tokens

POS tagging

- Different POS tags have different prior probability
- Most frequent class baseline - always assign apriori most frequent tag.
  - on WSJ corpus gives accuracy around 92%
  - more advanced classifiers give accuracy 97%.
- Given numbers - for English language
  - for other languages - lower
  - why English is so simple?