Morphological analysis

Distance

«Высшая школа экономики»

1.10

Outline

Morphological analysis

Морфологический анализ - процесс выявления структуры слов.

- Information retrieval (phone = phones $\neq$ phoned)
- Language modeling (scrutinize)
- Machine Translation (noun ➔ noun)

Morphological analysis

Word Formation

- Inflection (Lemmatization, Stemming)
  Derivation = prefix + stem + affix
  - friend + -ly = friendly
  - un- + do = undo
- Compounding = stem + stem
  järn(iron) + väg(road) = järnväg(railway)

Morphological analysis

token = lemma + POS + grammar feature

- singular vs. plural
- past, simple, future
- etc.

POS-tagging

- **Lexical Based Methods** — Assigns the POS tag the most frequently occurring with a word in the training corpus
- **Rule-Based Methods** — Assigns POS tags based on rules. For example, we can have a rule that says, words ending with “ed” or “ing” must be assigned to a verb. Rule-Based Techniques can be used along with Lexical Based approaches to allow POS Tagging of words that are not present in the training corpus but are there in the testing data.

POS-tagging

- **Probabilistic Methods** — This method assigns the POS tags based on the probability of a particular tag sequence occurring. Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) are probabilistic approaches to assign a POS Tag.
- **Deep Learning Methods** — Recurrent Neural Networks can also be used for POS tagging.

POS-tagging

- "I LOVE you, honey" vs. "Lets make LOVE, honey"
- Text to Speech Conversion

They *refuse* to permit us to obtain the *refuse* permit.

- refUSE (/rə'fy$\overline{oo}$z/) V
- refUSE (/refy,$\overline{oo}$s/) N

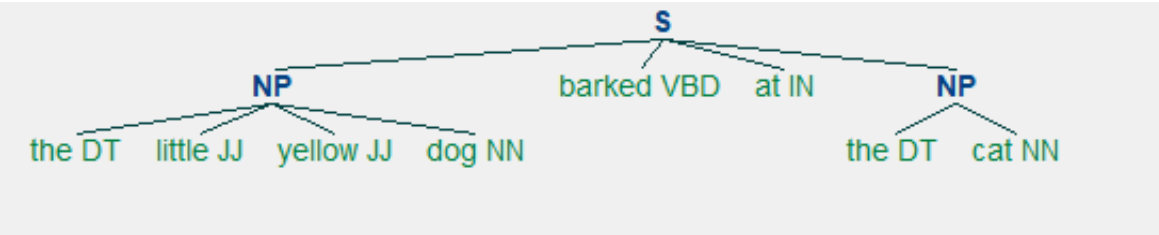
POS-tagging

- Noun (N)- Daniel, London, table, dog, teacher, pen, city, happiness, hope
- Verb (V)- go, speak, run, eat, play, live, walk, have, like, are, is
- Adjective(ADJ)- big, happy, green, young, fun, crazy, three
- Adverb(ADV)- slowly, quietly, very, always, never, too, well, tomorrow
- Preposition (P)- at, on, in, from, with, near, between, about, under
- Conjunction (CON)- and, or, but, because, so, yet, unless, since, if
- Pronoun(PRO)- I, you, we, they, he, she, it, me, us, them, him, her, this
- Interjection (INT)- Ouch! Wow! Great! Help! Oh! Hey! Hi!

Chunking

the little yellow dog barked at the cat

REGEX = NP: <DT>?<JJ>*<NN>



check NLTK page

Lexical distances

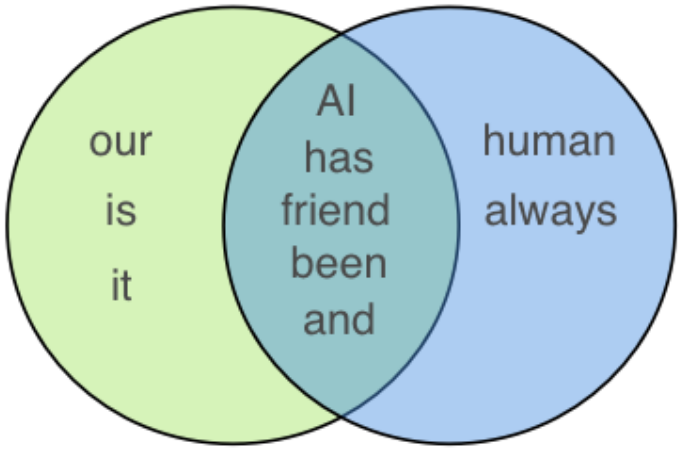
Levenshtein distance

- sitten ➔ sittin
- kitten ➔ sitten
- sittin ➔ sitting

Lexical distances

Jaccard Similarity

- Sentence 1: AI is our friend and it has been friendly
- Sentence 2: AI and humans have always been friendly



Venn Diagram of the two sentences for Jaccard similarity

Lexical distances

Cosine distance

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Cosine Similarity calculation for two vectors A and B [source]