University of BRISTOL

4.1 Part of Speech Tagging

Edwin Simpson

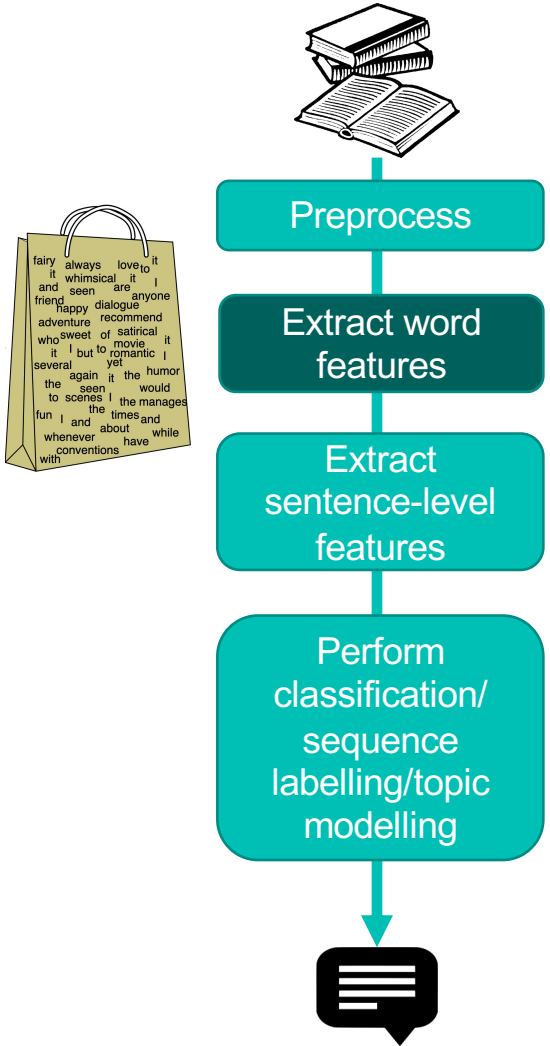
Department of Computer Science,

University of Bristol, UK.

bristol.ac.uk

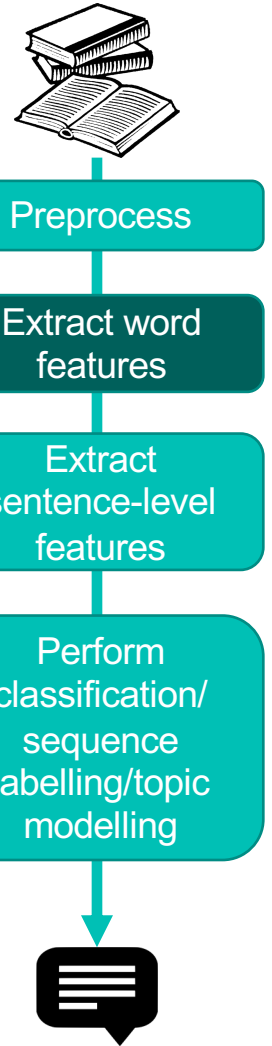
Previously…

- Bag-of-words: ignores structure and treats words as independent features.
- Vector representations of documents.
- What we're missing?
  - Disambiguation of words
  - Features for individual words
  - Syntactic structure of a sentence: how words relate to each other



Parts of Speech (POS)

- Nouns, verbs, adjectives, pronouns, prepositions, …
- Important information for downstream tasks:
  - POS help identify which words relate to each other, e.g., the subject of a verb
  - Information extraction – labelling entities and events, identifying their relations from verb phrases.
  - Sentiment analysis -- roles of adjectives in expressing sentiment are very different to verbs.
- Syntactic rather than semantic: they concern how words can be used in a sentence.



Open Classes

Chapter 8, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2019).

Nouns	People, places, things, ideas	*Obama, cat, room, London, consideration*
Verbs	Actions, processes	*Do, wait, seem, walk*
Adjectives	Descriptors	*Red, happy, unobtainable*
Adverbs	Modifiers	*Quickly, always, here,*

Closed Classes

Chapter 8, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2019).

Highly language-specific. Most important closed classes in English:

Prepositions	*on, under, over, near, by, at, from, to, with*
Particles	*up, down, on, off, in, out, at, by*
Determiners	*a, an, the*
Conjunctions	*and, but, or, as, if, when*
Pronouns	*she, who, I, others*
Auxiliary verbs	*can, may, should, are*
Numerals	*one, two, three, first, second, third*

POS Tagsets

- Classes can be defined in various ways, so use standard **tagsets.**
- **Penn Treebank** [1] for English with 45 tags:
  - Brown, WSJ and Switchboard corpora;
  - Hand-corrected tags to use for training POS taggers.
- **Universal dependencies** [2] provides 16 tags for any language:
  - Mappings for standard tagsets in at least 23 languages.
- Extra tags needed for social media [3].

[1] Marcus, M. P., et al.(1993). Building a large annotated corpus of English: The Penn treebank. Computational Linguistics.
[2] Nivre, J., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. LREC.
[3] Gimpel, K., et al. (2011). Part-of-speech tagging for Twitter: anno- tation, features, and experiments. ACL

POS Tagging in English

Section 8.3, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2019).

- Task: assign a POS tag to each word in a sentence.
- Requires disambiguation:
  - 14-15% of English vocab. is word types with multiple possible POS tags;
  - These words make up 55-67% tokens in a document!
- Most frequent class baseline: 92% accuracy on WSJ.

*earnings growth took a **back/JJ** seat* ← **adjective**

*a small building in the **back/NN*** ← **singular or mass noun**

*a clear majority of senators **back/VBP** the bill* ← **non-3rd person singular verb**

Morphologically Rich Languages

- E.g., Czech, Hungarian, Turkish
- Much more information than English in the **morphology** of the word
- Information like case, gender, person is important for downstream tasks like resolving references to an entity.
- Use a sequence of tags for each word

*Yerdeki **izin** temizlenmesi gerek.* → The trace on the floor should be cleaned.
POS tag of *izin* = iz+`Noun+A3sg+Pnon+Gen` [1]

Section 8.7, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2019).
[1] Hakkani-Tür, D., et al. (2002). Statistical morphological disambiguation for agglutinative languages. Journal of Computers and Humanities

POS Tagging in Chinese

- Very short words compared to English;
- Tokenisation is difficult (see video 1.3) so POS tagging and tokenisation may be done at the same time;
- Compounding produces many unknown common nouns and verbs.
- To perform POS tagging, we use various features:
  - Prefixes and suffixes
  - Elements of the characters such as radicals.

POS Tagging as Sequence Labelling

- Task: assign class labels to tokens in a sequence.
- When tagging a word, consider its neighbours in the sequence to disambiguate the tag.
- Sequence labelling methods achieve ~97% accuracy on WSJ.
  - Hidden Markov models (HMMs)
  - Conditional random fields (CRFs)
  - Recurrent neural networks (RNNs)
- Sequence labellers have broad applications in text analytics, e.g.:
  - Information extraction: identifying words that refer to people or events;
  - Question answering: find a spans of text that answer a user's question.

Summary

- Syntactic features are useful for many text analytics tasks, such as sentiment analysis and information extraction.
- Part-of-speech tags are a very useful type of syntatic feature that indicate a word's usage in a sentence.
- POS tags vary greatly between languages but most include nouns, verbs and adjectives, while some languages may require a complex sequence of tags to label a single word.
- POS tagging is itself a sequence labelling task, as the tag of an individual word depends on the previous and next words.