Tagging: Classification in Context

CS 490A, Fall 2020

10/1/2020

Applications of Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs490a_f20/

Brendan O'Connor

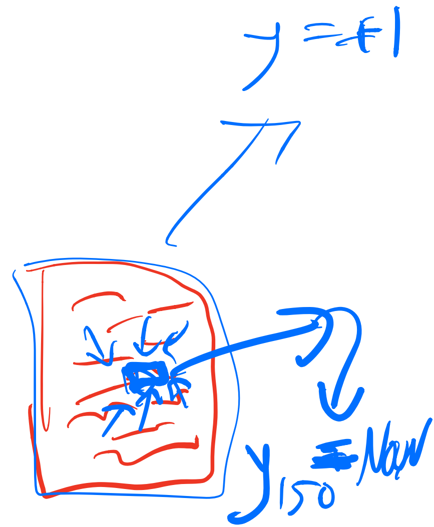
College of Information and Computer Sciences

University of Massachusetts Amherst

In-text classification

Doc Classif.

- Previous: Text Classification → Doc, Sentence
  - Input: a whole text
  - Output: → Category



- Let's move to classifying **within the text!**
  - Tasks you can do yourself, with the right heuristics or logistic regression features (or other NLP models)
  - Do it with a pretrained, off-the-shelf system as part of a larger system, especially for syntactic/semantic linguistic analyses

- Tagging
  - Input: seq. of tokens
  - Output: categ. per token

I feel great.

NOUN VERB ADJ

- Span classification
  - Input: span (subseq) of tokens
  - Output: categ.

0 1 2 3 4 5

Where's the Sc. Sci. bldg?

ORG-NAME

span: [2, 4)

- Relation classification
  - Input:
  - Output:

Targeted sentiment analysis

POS NEG

(4.2) a. The vodka was good, but the meat was rotten
b. Go to Heaven for the climate, Hell for the company. –*Mark Twain*

Meat
Vodka was

fairly rotten
pretty rotten

Hard problem!

Word sense disambiguation

LEADER WEAPONS

(4.3) a. Iraqi head seeks arms

b. Prostitutes appeal to Pope

c. Drunk gets nine years in violin case[2]

Context features!

Iraqi head seeks arms

BODY PART BODY-PART

Probabilistic

Part of speech tags

VERB
NOUN

- Syntax = how words compose to form larger meaning-bearing units
- POS = syntactic categories for words
  - You could substitute words within a class and have a syntactically valid sentence.
  - Give information how words can combine.

NOUN

  - I saw the dog
  - I saw the cat
  - I saw the {table, sky, dream, school, anger, ...}

Schoolhouse Rock: Conjunction Junction
https://www.youtube.com/watch?v=ODGA7ssL-6g&index=1&list=PL6795522EAD6CE2F7

Part of speech tagging

PRO VERB NOUN

- I saw the fire today

- Fire!

ambig!

Noun
Verb

Open vs closed classes

- Open class (lexical) words
  - Nouns
    - Proper: *IBM*, *Italy*
    - Common: *cat / cats*, *snow*
  - Verbs
    - Main: *see*, *registered*
  - Adjectives: *old older oldest*
  - Adverbs: *slowly*
  - Numbers: *122,312*, *one*
  - … more
- Closed class (functional)
  - Determiners: *the some*
  - Conjunctions: *and or*
  - Pronouns: *he its*
  - Verbs
    - Modals: *can*, *had*
  - Prepositions: *to with*
  - Particles: *off up*
  - Interjections: *Ow Eh*
  - … more

slide credit: Chris Manning

Why do we want POS?

- Useful for many syntactic and other NLP tasks.
  - Phrase identification ("chunking")
  - Named entity recognition (names = proper nouns... or are they?)
  - Syntactic/semantic dependency parsing
  - Sentiment
- Either as features or heuristic filtering
- Esp. useful when not much training data

POS patterns: simple noun phrases

- Quick and dirty noun phrase identification
  http://brenocon.com/JustesonKatz1995.pdf
  http://brenocon.com/handler2016phrases.pdf

Grammatical structure: Candidate strings are those multi-word noun phrases that are specified by the regular expression $((A \mid N)^+ \mid ((A \mid N)^*(NP)^?)(A \mid N)^*)N$,

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

Table 5.2 Part of speech tag patterns for collocation filtering. These patterns were used by Justeson and Katz to identify likely collocations among frequently occurring word sequences.

POS patterns: sentiment

- Turney (2002): identify bigram phrases, from unlabeled corpus, useful for sentiment analysis.

Table 1. Patterns of tags for extracting two-word phrases from reviews.

	First Word	Second Word	Third Word (Not Extracted)
1.	JJ	NN or NNS	anything
2.	RB, RBR, or RBS	JJ	not NN nor NNS
3.	JJ	JJ	not NN nor NNS
4.	NN or NNS	JJ	not NN nor NNS
5.	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Table 2. An example of the processing of a review that the author has classified as *recommended*.[6]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732

(plus co-occurrence information)

Named entity recognition

"The The a

SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

ORG MISC

Figure 1: Example illustrating challenges in NER.

PERsons

LOCation

ORGanization

MISC

[Ratinov and Roth 2009]

Useful features for a tagger

- Key sources of information:
  - 1. The word itself
  - 2. Word-internal characters
  - 3. Nearby words in a *context window*
    - **Context window features are used for ALL tagging tasks!**

-ed

Features for NER/POS

- Word-based features
  - Word itself
  - Word shape
  - Contextual variants: versions of these at position t-1, t-2, t-3 … t+1, t+2, t+3 …

Aa AA

"bwat-at_t_-1"

"bwt-at-t_+1"

- External lexical knowledge
  - Gazetteer features: Does word/phrase occur in a list of known names?
  - Other hand-built lexicons
  - Word embeddings (next week)

Gazetteers example

Wikepedia

1)**People**: *people, births, deaths*. Extracts 494,699 Wikipedia titles and 382,336 redirect links. 2)**Organizations**: *cooperatives, federations, teams, clubs, departments, organizations, organisations, banks, legislatures, record labels, constructors, manufacturers, ministries, ministers, military units, military formations, universities, radio stations, newspapers, broadcasters, political parties, television networks, companies, businesses, agencies*. Extracts 124,403 titles and 130,588 redirects. 3)**Locations**: *airports, districts, regions, countries, areas, lakes, seas, oceans, towns, villages, parks, bays, bases, cities, landmarks, rivers, valleys, deserts, locations, places, neighborhoods*. Extracts 211,872 titles and 194,049 redirects. 4)**Named Objects**: *aircraft, spacecraft, tanks, rifles, weapons, ships, firearms, automobiles, computers, boats*. Extracts 28,739 titles and 31,389 redirects. 5)**Art Work**: *novels, books, paintings, operas, plays*. Extracts 39,800 titles and 34037 redirects. 6)**Films**: *films, telenovelas, shows, musicals*. Extracts 50,454 titles and 49,252 redirects. 7)**Songs**: *songs, singles, albums*. Extracts 109,645 titles and 67,473 redirects. 8)**Events**: *playoffs, championships, races, competitions, battles*. Extracts 20,176 titles and 15,182 redirects.

[Ratinov and Roth 2009]

What is a baseline

- Most frequent class

- Someone else's pretrained model