NATURAL LANGUAGE PROCESSING

Introduction

TOPICS

- **Natural Language Processing**
- Regular Expression
- Text Tokenization
- Part-Of-Speech Tagging

LANGUAGE IS AMBIGUOUS

- Example: I saw a man on a hill with a telescope.
- Machine's Interpretations:
  - There is a man on a hill, I saw him with a telescope.
  - There is a man on a hill, he has a telescope. And I saw him.
  - There is a man and he is on the hill. There is a telescope also on the same hill.
  - I am on a hill, and I saw a man with a telescope.

Yanjun Li

3

NLP IS HARD

- NLP is a field at the intersection of ***Computer Science, Artificial Intelligence,*** and ***Linguistics.***
- Goal of NLP : for computers to process or "understand" natural language in order to perform tasks that are useful, such as Language Translation, Question Answering, Classifying, and so on.
- Humans are the 'gold standard' for 'intelligence' and even we make mistakes when it comes to understanding language.
- Interpretation of language is colored by feelings, past experiences, context, …

Yanjun Li

4

WHAT NLP CAN DO

- Online advertisement matching
- Online search
- Automated translation
- Sentiment analysis for marketing
- Chatbots/Dialog agents
  - Automating customer support
  - Controlling devices
  - Ordering goods

TOPICS

- Natural Language Processing
- **Regular Expression**
- Text Tokenization
- Part-Of-Speech Tagging

REGULAR EXPRESSION

- A formal language for specifying text strings, an important theoretical tool throughout computer science and linguistics.
- A regular expression is an algebraic notation for characterizing a set of strings.
  - A string is any sequence of alphanumeric characters.
- Regular expression search requires a **pattern** that we want to search for and a **corpus** of texts to search through.

Yanjun Li

7

REGULAR EXPRESSIONS: DISJUNCTIONS

- **Match one of letters inside square brackets [ ]**

Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my beans were impatient
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Yanjun Li

8

REGULAR EXPRESSIONS: MORE DISJUNCTION

- *Woodchucks* is another name for *groundhog*!
- The pipe | for disjunction (one of words)

Pattern	Matches
groundhog|woodchuck	groundhog woodchuck
yours|mine	yours Mine
a|b|c	= [abc]
[gG]roundhog|[Ww]oodchuck	groundhog Groundhog woodchuck Woodchuck

Yanjun Li

9

REGULAR EXPRESSIONS: NEGATION IN DISJUNCTION

- **Negations** **[^Ss] (not any of letters)**
  - Caret means negation only when showing as the first symbol in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason
a^b	The pattern 'a caret b'	Look up a^b now
	Not a negation	

Yanjun Li

10

REGULAR EXPRESSIONS: ? * + .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	Zero or more of previous char	oh! ooh! oooh! ooooh!
[ab]*	Zero or more of a or b	aaa bb ababa
o+h!	One or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
[0-9]+	A sequence of digits	
beg.n	Wildcard, any char Except a newline	begin begun begun beg3n

Yanjun Li

11

REGULAR EXPRESSIONS: ANCHORS ^ $

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
end\.$	The end.
end.$	The end? The end!

Yanjun Li

12

REGULAR EXPRESSIONS: OPERATOR PRECEDENCE

- Python nltk package

Pattern	Matches
a(b|c)+	Parentheses that indicate the scope of the operators
{n}	Exactly n repeats where n is a non-negative integer
{n,}	At least n repeats
{, n}	No more than n repeats
{m,n}	At least m and no more than n repeats

EXAMPLE

- Find me all instances of the word "the" in a text.

`the` Misses capitalized examples

`[tT]he` Incorrectly returns `other` or `theology`

`[^a-zA-Z][tT]he[^a-zA-Z]`

Misses the as the first word in a line or the last word in a line

`(^|[^a-zA-Z])[tT]he([^a-zA-Z]|$)`

SUMMARY

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing tool.
- For many hard tasks, we use machine learning classifiers
  - But regular expressions are used as features in the classifiers
  - Can be very useful in capturing generalizations

TOPICS

- Natural Language Processing
- Regular Expression
- **Text Tokenization**
- Part-Of-Speech Tagging

HOW MANY WORDS?

- "Seuss's cat in the hat is different from other cats! "
  - **Lemma**: same stem, part of speech, rough word sense
    - cat and cats = same lemma
  - **Wordform**: the full inflected surface form
    - cat and cats = different wordforms

Yanjun Li

17

HOW MANY WORDS?

They picniced by the pool, then lay back on the grass and looked at the stars.

- **Type**: an element of the vocabulary, a.k.a. unique word term.
- **Token**: an instance of that type in running text.
- How many?
  - 16 tokens (or 18)
  - 14 types (or 16)

Yanjun Li

18

HOW MANY WORDS?

N = number of tokens

V = vocabulary = set of types

|*V*| is the size of the vocabulary

Church and Gale (1990): $|V| > O(N^{½})$

	Tokens = N	Types = |V|
Switchboard phone conversations	2,400,000	20,000
Shakespeare	884,000	31,000
Google N-grams	1,000, 000,000,000	13,000,000

CORPORA

- When a computation model is developed for language processing, it is important to consider who produced, in what context, for what purpose.
  - Language with dialect : Standard American English
  - Genre
  - Time

TEXT NORMALIZATION

- Every NLP task needs to do text normalization:
  1. Segmenting/tokenizing words in running text
  2. Normalizing word formats
  3. Segmenting sentences in running text

Yanjun Li

21

WORD TOKENIZATION

- Tokenization: The task of segmenting running text into words
- The standard method for tokenization is to use deterministic algorithms based on ***regular expressions*** compiled into very efficient finite state automata.
- Penn Treebank Tokenization Standard

Yanjun Li

22

ISSUES IN TOKENIZATION

- `Finland's capital` → `Finland Finlands Finland's ?`
- `what're, I'm, isn't` → `What are, I am, is not`
- `Hewlett-Packard` → `Hewlett Packard ?`
- `state-of-the-art` → `state of the art ?`
- `Lowercase` → `lower-case lowercase lower case ?`
- `San Francisco` → one token or two?
- m.p.h., PhD. → ??

Yanjun Li

23

WORD NORMALIZATION

- Normalization: The task of putting words/tokens in a standard format.
  - Information Retrieval: indexed text & query terms must have same form.
    - We want to match ***U.S.A.*** and ***USA***
- We implicitly define equivalence classes of terms
  - e.g., deleting periods in a term
- Alternative: asymmetric expansion:
  - Enter: ***window*** Search: ***window, windows***
  - Enter: ***windows*** Search: ***Windows, windows, window***
  - Enter: ***Windows*** Search: ***Windows***
- Potentially more powerful, but less efficient

Yanjun Li

24

LEMMATIZATION

- Reduce inflections or variant forms to base form
  - *am, are, is* $\rightarrow$ *be*
  - *car, cars, car's, cars'* $\rightarrow$ *car*
- *the boy's cars are different colors* $\rightarrow$ *the boy car be different color*
- Lemmatization: have to find correct dictionary headword form
- Machine translation
  - Spanish quiero ('I want'), quieres ('you want') same lemma as querer 'want'

Yanjun Li

25

STOP WORDS

- Some extremely common words appear to be of little value in performing certain tasks.
- A stop word list is created based on collection frequency.
- Example: "a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the….."
- In practice, stop words are removed from the data.

Yanjun Li

26

MORPHOLOGY

- **Morphemes**:
  - The small meaningful units that make up words
  - **Stems**: The core meaning-bearing units
  - **Affixes**: Bits and pieces that adhere to stems
    - Often with grammatical functions

Yanjun Li

27

STEMMING

- Reduce terms to their stems in information retrieval
- *Stemming* is crude chopping of affixes
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

→

for exampl compress and compress ar both accept as equival to compress

Yanjun Li

28

PORTER'S ALGORITHM THE MOST COMMON ENGLISH STEMMER

Step 1a

```
sses → ss    caresses → caress
ies  → i     ponies   → poni
ss   → ss    caress   → caress
s    → ø     cats     → cat
```

Step 1b

```
(*v*)ing → ø  walking   → walk
              sing      → sing
(*v*)ed  → ø  plastered → plaster
…
```

Step 2 (for long stems)

```
ational→ ate  relational→ relate
izer→ ize     digitizer → digitize
ator→ ate     operator  → operate
…
```

Step 3 (for longer stems)

```
al    → ø  revival    → reviv
able  → ø  adjustable → adjust
ate   → ø  activate   → activ
…
```

Yanjun Li

29

SENTENCE SEGMENTATION

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
  - Sentence boundary
  - Abbreviations like Inc. or Dr.
  - Numbers like .02% or 4.3
- Build a binary classifier
  - Looks at a "."
  - Decides EndOfSentence/NotEndOfSentence
  - Classifiers: hand-written rules, regular expressions, or machine-learning

Yanjun Li

30

TOPICS

- Natural Language Processing
- Regular Expression
- Text Tokenization
- **Part-Of-Speech Tagging**

PART-OF-SPEECH TAGGING

- 8 (ish) traditional parts of speech
  - Large amount of information about a word and its neighbors.
  - Noun, verb, adjective, preposition, adverb, article, interjection, pronoun, conjunction, etc
  - Called: *parts-of-speech, lexical categories, word classes, morphological classes, lexical tags…*

POS EXAMPLES

N	noun	*chair, bandwidth, pacing*
V	verb	*study, debate, munch*
ADJ	adjective	*purple, tall, ridiculous*
ADV	adverb	*unfortunately, slowly*
P	preposition	*of, by, to*
PRO	pronoun	*I, me, mine*
DET	determiner	*the, a, that, those*

Yanjun Li

33

POS TAGGING

- The process of assigning a part-of-speech or lexical class marker to each word in a collection.

WORD	TAG
the	**DET**
koala	**N**
put	**V**
the	**DET**
keys	**N**
on	**P**
the	**DET**
table	**N**

Yanjun Li

34

WHY IS POS TAGGING USEFUL?

- First step of a vast number of practical tasks
- Parsing
  - Need to know if a word is an N or V before you can parse
- Information extraction
  - Finding names, relations, etc.
- Machine Translation

OPEN AND CLOSED CLASSES

- Closed class: a small fixed membership
  - Prepositions: of, in, by, …
  - Auxiliaries: may, can, will had, been, …
  - Pronouns: I, you, she, mine, his, them, …
  - Usually function words (short common words which play a role in grammar)
- Open class: new ones can be created all the time
  - English has 4: Nouns, Verbs, Adjectives, Adverbs
  - Many languages have these 4, but not all!

POS TAGGING
CHOOSING A TAGSET

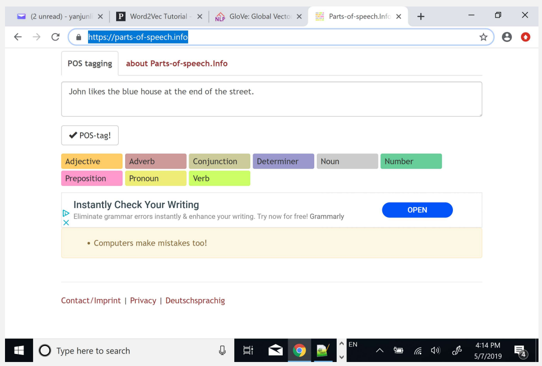
- There are so many parts of speech, potential distinctions we can draw
- To do POS tagging, we need to choose a standard set of tags to work with
- Could pick very coarse tagsets
  - N, V, Adj, Adv.
- More commonly used set is finer grained, the "Penn TreeBank tagset", 45 tags (*pos_tag* in **nltk**)
  - PRP$, WRB, WP$, VBG
- Even more fine-grained tagsets exist

POS TAGGING

- Words often have more than one POS: *back*
  - The *back* door = JJ (adjective)
  - On my *back* = NN (noun)
  - Win the voters *back* = RB (adverb)
  - Promised to *back* the bill = VB (verb base form)
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

These examples from Dekang Lin

ONLINE RESOURCE



TWO METHODS FOR POS TAGGING

1. Rule-based tagging
   - EngCG (ENGTWOL)
2. Stochastic - Probabilistic sequence models
   - HMM (Hidden Markov Model) tagging
   - Recurrent Neural Network

RULE-BASED TAGGING

- Start with a dictionary
- Assign all possible tags to words from the dictionary
- Write rules by hand to selectively remove tags
- Leaving the correct tag for each word.

HIDDEN MARKOV MODEL TAGGING

- Using an HMM to do POS tagging is a special case of *Bayesian inference*
  - Foundational work in computational linguistics
  - Bledsoe 1959: OCR
  - Mosteller and Wallace 1964: authorship identification
- It is also related to the "noisy channel" model that's the basis for Automatic Speech Recognition, Optical Character Recognition and Machine Translation.

EVALUATION

- Overall error rate with respect to a gold-standard test set.
- Error rates on particular tags
- Error rates on particular words
- Tag confusions...