Natural Language Processing: **Word-level Processing**

Hutchatai Chanlekha

Department of Computer Engineering, Faculty of Engineering
Kasetsart University

Let’s discuss the output from homework

- What are the top 20 words?
- What does the information from frequency of frequency table tell you?

WORD STATISTICS

Word counts (cont.)

- How many words are there in text?
  - How many *word tokens* there are?
  - How many *different words* (i.e. *word types*) appear in the text?
  - Example: Tom Sawyer
    - 71370 word tokens
    - 8018 distinct words
    - Ratio of tokens to types
      - average frequency with which each type is used
      - 71370/8018 = 8.9

Word count (cont.)

- Ratio of tokens to types statistic tell us that how often words in the corpus occur "on average"
- BUT!!!
  - Word types have a very uneven distribution
    - From very rare to very common
  - From the table
    - 49.8% of word types occur only once.
    - Over 90% of word types occur 10 times or less.
- It is hard to predict much about the behavior of words that we never or barely ever observed in the corpus

Word frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
>100	102

F of F of word types in Tom Sawyer*

* Taken from 'Statistical Foundation of Natural Language Processing'

Zipf's laws

- Count up how often each word occurs in a large corpus
- Then list the words in order of their frequency of occurrence

Word	Freq (f)	Rank (r)	f*r
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
oh	116	90	10440
two	104	100	10400

Word	Freq (f)	Rank (r)	f*r
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
could	2	4000	8000
applausive	1	8000	8000

Based on Tom Sawyer

Zipf's laws (cont.)

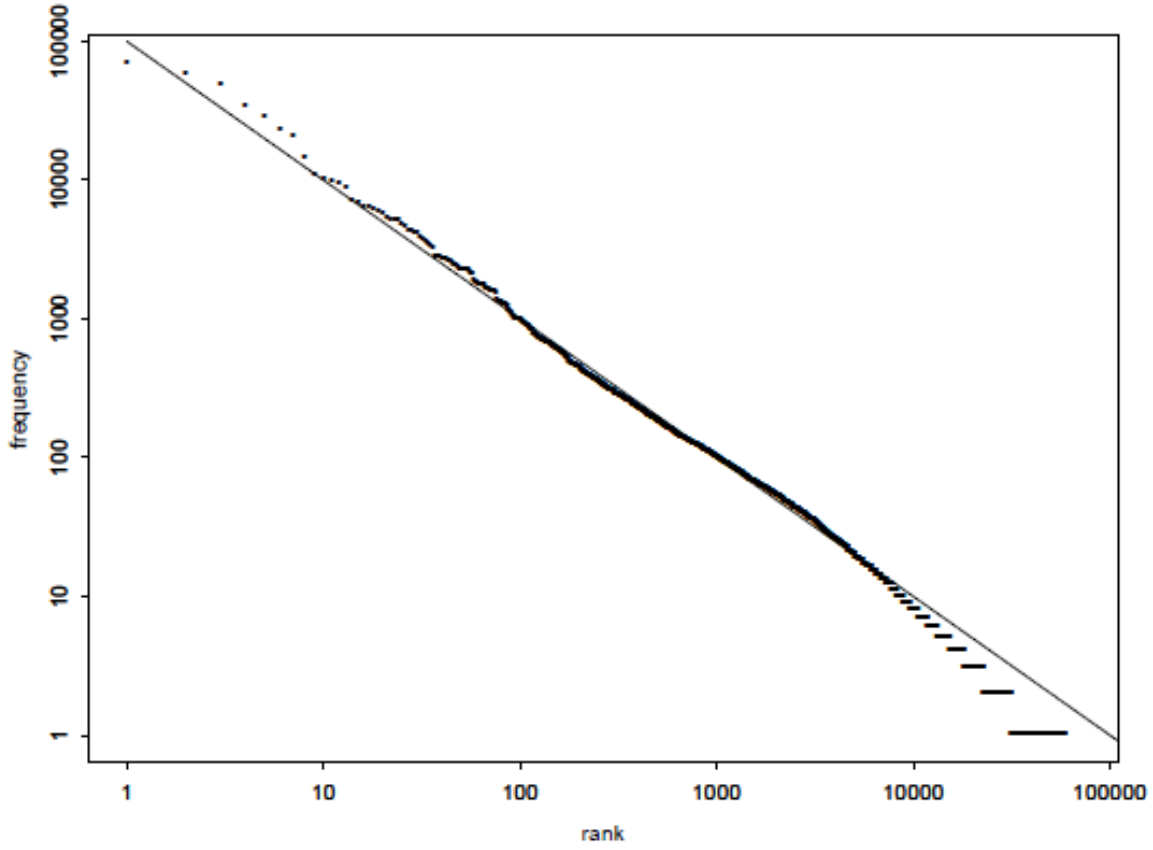
- Zipf's law says that $f \propto \frac{1}{r}$

  (there is a constant k such that: f*r = k)

- From previous table:
  - Zipt's law is shown to approximately hold.
    - Except the first three highest frequency words
    - Bulge a little for words of rank around 100
  - Nevertheless, Zipf's law is useful as a rough description of the frequency distribution of words in human languages:
    - A few very common words
    - A middling number of medium frequency words
    - Many low frequency words

Zipf's laws (cont.)

- Mandelbrot (1954) studied Zipf's law and noted that if plot rank-frequency on doubly logarithmic axes, the line is often a bad fit, especially for low and high ranks.



Brown corpus

Zipf's laws (cont.)

- Other laws
  - Zipf argues that conservation of speaker effort would prefer there to be only one word with all meanings, while conservation of hearer effort would prefer each meaning to be expressed by a different word.
    - Relation between number of meanings of a word

$$m \propto \sqrt{f} \qquad m \propto \frac{1}{\sqrt{r}}$$

  - Tendency of content words to clump
    - Measure the number of lines or pages between each occurrence of the word
    - Then calculate frequency F of different interval sizes I.
    - Based on studied corpus, for words of frequency at most 24

$$F \propto I^{-\rho}$$

Number of intervals of a certain size was inversely related to the interval size

    - Means that most of the time, content words occur near another occurrence of the same word.
  - Inverse relationship between the frequency of words and their length

Zipf's laws (cont.)

- Most of the languages seem to follow Zipf's law
  - less valuable as a characterization of language
- Highlight the reason that makes frequency-based approaches to language hard
  - Almost all words are rare

WORD PROCESSING

การประมวลผลคำ: ความครอบคลุมของคำศัพท์

- หากคลังคำศัพท์ไม่ครอบคลุม ระบบอาจไม่สามารถทำงานได้อย่างถูกต้อง
- การสร้างฐานคำศัพท์ซึ่งประกอบด้วยคำศัพท์ และข้อมูลในคำศัพท์ให้สมบูรณ์ ไม่ใช่เรื่องง่าย
  - คน vs เครื่องคอมพิวเตอร์
  - คำใหม่เกิดขึ้นตลอดเวลา

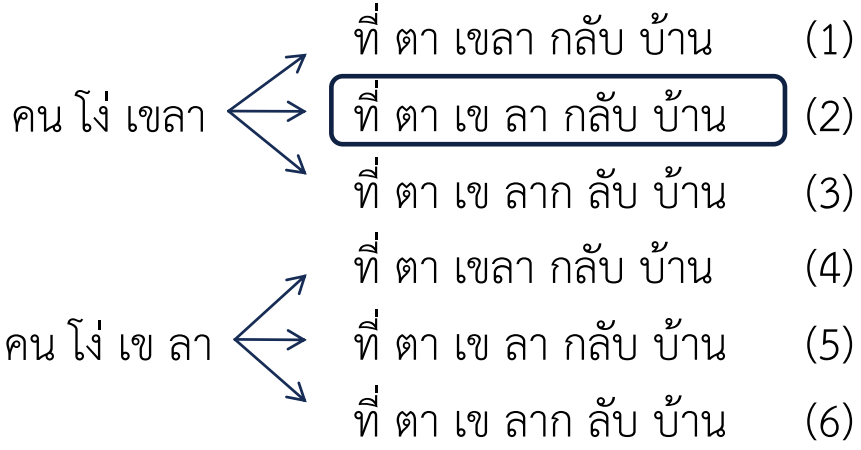
Word-level Ambiguity

- 1 คำ หลายความหมาย หลาย POS

“ตั้ง”	หน้าที่	ความหมาย
1. น้องตั้งแจกันไว้บนโต๊ะ	สกรรมกริยา	วาง
2. โรงแรมตั้งมานานแล้ว	อกรรมกริยา	ดำรงอยู่
3. เขาหอบหนังสือมาสองตั้ง	ลักษณนาม	ของที่วางซ้อนๆ กัน
“ไข่”	**หน้าที่**	**ความหมาย**
1. ไข่แตกแล้ว	คำนาม	อาหารชนิดหนึ่ง
2. แม่ไก่ไข่แล้ว	อกรรมกริยา	อาการตกฟอง
“ขัน”	**หน้าที่**	**ความหมาย**
1. เอาขันมาตักน้ำ	คำนาม	ภาชนะชนิดหนึ่ง
2. เขาพูดขันเหลือเกิน	กริยาวิเศษณ์	น่าหัวเราะ
3. อย่าลืมขันน็อตให้แน่น	สกรรมกริยา	ปิดให้แน่น

ความคลุมเครือขอบเขตของคำ

- ภาษาไทย ไม่มีการใช้ช่องว่างเพื่อแบ่งขอบเขตระหว่างคำ
- ตัดคำผิด → ส่งผลต่อความถูกต้องของ POS tagging, parsing, semantic understanding, etc.
- Example: คนโง่เขลาที่ตาเขลากลับบ้าน



Word Processing

- Word segmentation
- Stemming/Lemmatization
- Word formation
- POS-tagging
- …

Word segmentation

- Approaches
  - Rule based
    - Spelling rule
  - Dictionary based
    - Longest matching
    - Shortest matching
    - Maximum matching
  - Statistical based
    - N-gram
    - Machine learning, such as HMM, Winnow, Deep Learning, etc.

Look at the example …

There are cats in his house that are starving.

there was a cat in her room that was starved.

- Do you think these two sentences are about the same topic?

What we see …
- “there” and “There” are the same
- “are” and “was” are both verb-to-be
- “cats” is plural form of “cat”
- “his” and “her” are both possessive adj.
- “starving” and “starved” are inflection of the same word

All words are almost the same!

What computer sees …
- “there” and “There” are different
- “are” and “was” are different
- “cat” and “cats” are different
- “his” and “her” are different
- “house” and “room” are living place
- “starving” and starved” are different

There are 6 words that are different!

- How to make computers know these words are similar?

Common Pre-processing in NLP Application

- Stemming & Lemmatization
  - Finding root word or lexeme
  - What are the differences between stemming and lemmatization?

Sentence I	**It**	**was**	**very**	**fantastic**	**experience**
Lemma	it	be	very	fantastic	experience
Stemming	it	wa	veri	fantast	experi

Sentence II	**John**	**sent**	**an**	**email**	**to**	**Mary**
Lemma	John	send	a	email	to	Mary
Stemming	John	sent	an	email	to	Mari

SentenceIII	**He**	**didn't**		**get**	**a**	**reply**
Lemma	he	do	not	get	a	reply
Stemming	he	didn	t	get	a	repli

Stemming & Lemmatization

- Finding root word or lexeme
  - run, ran, running, runs → same lexeme
  - memory, memories, memorize, memorized, memorizing, memorizes, memorization, memorizer, memorizers, etc.
- Goal
  - Reduce inflectional forms and sometimes derivationally related forms of a word to a common base form

Stemming

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

- The most common algorithm for stemming English is *Porter's algorithm*
  - Has been shown to be empirically very effective

Example of stemmer algorithm

- Porter stemmer
  - Step 1: Gets rid of plurals and -ed or -ing suffixes
  - Step 2: Turns terminal y to i when there is another vowel in the stem
  - Step 3: Maps double suffixes to single ones: -ization, -ational, etc.
  - Step 4: Deals with suffixes, -full, -ness etc.
  - Step 5: Takes off -ant, -ence, etc.
  - Step 6: Removes a final -e

Ref: https://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/dan-porters.pdf

Questions!!

- The following pairs of words are stemmed to the same form by the Porter stemmer.
  - abandon/abandonment
  - marketing/markets
  - university/universe
  - volume/volumes
- Test: http://9ol.es/porter_js_demo.html
- Which pairs would you argue shouldn't be conflated? Give your reasoning.

http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Lemmatization

- *Lemmatizer* does full morphological analysis to accurately identify the lemma for each word.

Sentence I	**It**	**was**	**very**	**fantastic**	**experience**	
Lemma	it	be	very	fantastic	experience	
Stemming	it	wa	veri	fantast	experi	

Sentence II	**John**	**sent**	**an**	**email**	**to**	**Mary**
Lemma	John	send	a	email	to	Mary
Stemming	John	sent	an	email	to	Mary

SentenceIII	**He**	**didn't**		**get**	**a**	**reply**
Lemma	he	do	not	get	a	reply
Stemming	he	didn't		get	a	repli

Stemming by Porter stemming, Lemmatizing by Stanford Corenlp

Differences between stemming and lemmatization

Stemming

- Usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time
- Often includes the removal of derivational affixes.

Lemmatization

- Usually refers to doing things properly with the use of a vocabulary and morphological analysis of words
- Normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*

Test: https://corenlp.run/

More detail, see http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Question!!

- True or False??
  - Stemming increases the size of the vocabulary. 
  - In a document retrieval system, stemming rarely lowers precision. 
  - In a document retrieval system, stemming rarely lowers recall. 

http://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Word formation

- Word formation
  - In a form of inflection, derivation
    - Know root word, how to derive other possible words from root?
      - Run/ran/runs/running
      - Good/better/best
      - Complex/complexion/complexity
  - In a form of compound word
    - คำประสมที่ถูกต้องตามหลักภาษา
      - การ, ความ, ผู้, ชาว, นัก
        - การกิน VS ความกิน, ชาวนา VS นักนา, นักเรียน VS ชาวเรียน
      - คำประสมอื่นๆ
        - เก้าอี้, เก้าอี้ผ้าใบ, เก้าอี้ไม้, เก้าอี้โยก, เก้าอี้พลาสติก, ฯลฯ

Question!!

- What is the usefulness of stemming and lemmatization?
- What is the usefulness of word formation?

POS tagging

- Part-of-speech tagging (การกำกับชนิดของคำ)
  - Dictionary-based
  - Statistical-based or Machine learning

“ขัน”	หน้าที่	ความหมาย
1. เอาขันมาตักน้ำ	คำนาม	ภาชนะชนิดหนึ่ง
2. เขาพูดขันเหลือเกิน	กริยาวิเศษณ์	น่าหัวเราะ
3. อย่าลืมขันน็อตให้แน่น	สกรรมกริยา	ปิดให้แน่น

Test: https://corenlp.run/

POS tags in Penn Treebank

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural

Tag	Description
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

http://www.cis.upenn.edu/~treebank/

POS tags in Penn Treebank

Tag	Description
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Tag	Description
$	dollar
``	opening quotation mark
''	closing quotation mark
(	opening parenthesis
)	closing parenthesis
,	Comma
--	dash
.	sentence terminator
:	colon or ellipsis

More detail on part-of-speech tag set:
http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

http://www.cis.upenn.edu/~treebank/