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

Hutchatai Chanlekha

Department of Computer Engineering, Faculty of Engineering
Kasetsart University

Outline

- Words
- Words and Languages
- Word-level processing
  - Word segmentation
  - Lemmatization and Stemming
  - POS tagging
- N-gram model

Word

- คำ ประกอบขึ้นจากหน่วยคำ (morpheme)
- 1 คำ อาจประกอบขึ้นจาก 1 หน่วยคำ หรือหลายหน่วยคำ
  - “กิน”, “จมูก”, “สภาพ” -> 1 หน่วยคำ
  - “โรงเรียน”, “ชาวนา” -> 2 หน่วยคำ
- หน่วยคำ ต่างจากพยางค์
  - 1 หน่วยคำ อาจมี 1 พยางค์ หรือ หลายพยางค์

Part-Of-Speech (POS)

- ชนิดของคำ (part-of-speech)

ชนิดของคำ	ตัวอย่าง
นาม (noun)	เก้าอี้ ความงาม การกิน นักเรียน
กริยา (verb)	ไป วิ่ง หลับ
คุณศัพท์ (adjective)	(สี)ดำ สวย ตะกละ อ้วน
วิเศษณ์ (adverb)	บ่อยๆ ทันที อย่างกระตือรือร้น
บุพบท (preposition)	ใน ใต้ ของ
สันธาน (conjunction)	และ เพราะ ถ้า
สรรพนาม (pronoun)	หล่อน เธอ เขา มัน ฉัน
อุทาน (interjection)	อุ๊ย อ้าว

- กลุ่มคำที่มีหน้าที่เฉพาะ เช่น determiners ได้แก่ articles (a, an, the) และคำอื่นๆ (เช่น some, there) และ Particles เช่น up ใน give up, off ใน jump off

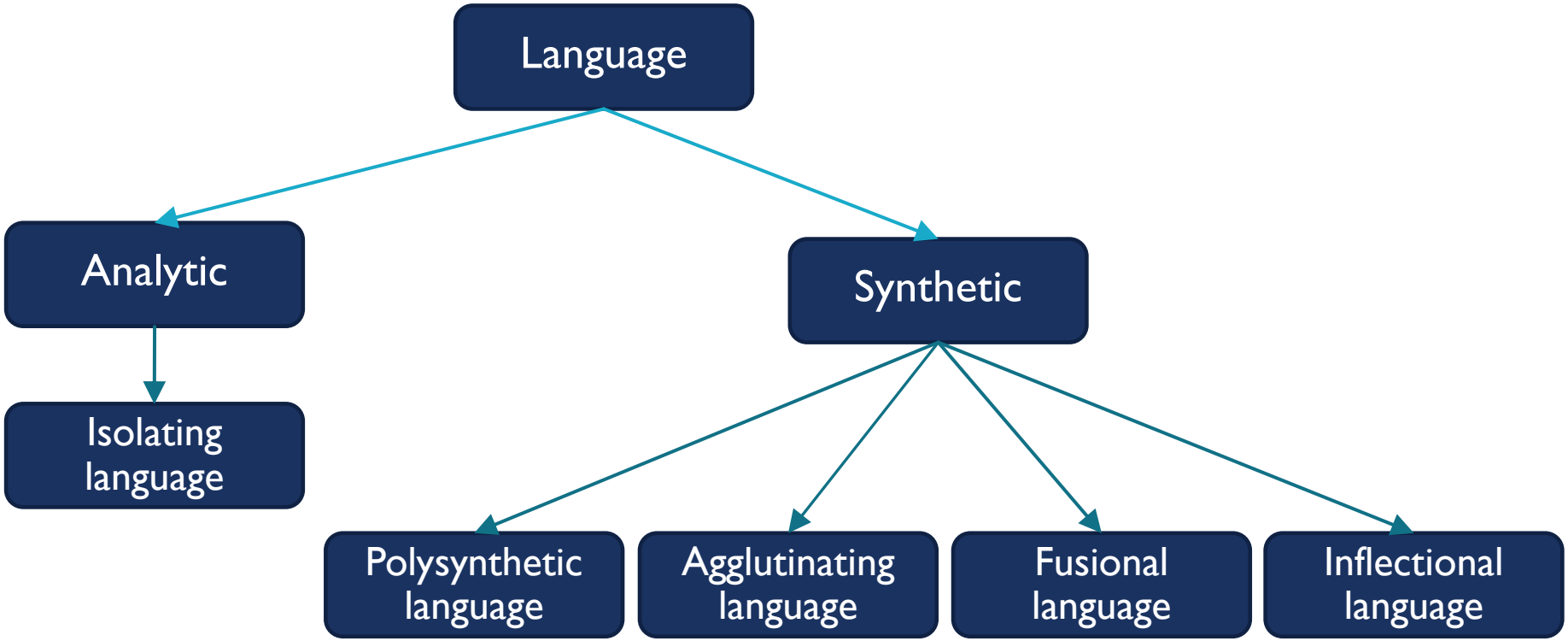
ข้อสนเทศของคำ

ข้อสนเทศ	ใช้กับ	แสดง
บุรุษ (person)	สรรพนาม, กริยา	แสดงให้เห็นถึงผู้พูด ผู้ฟัง บุคคลที่สามหรือที่สี่ที่กล่าวถึง (บุรุษที่ 1, บุรุษที่ 2, บุรุษที่ 3)
พจน์ (number)	สรรพนาม, นาม, กริยา	แสดงถึงจำนวนของส่วนประกอบ ได้แก่ เอกพจน์ พหูพจน์
เพศ (gender)	สรรพนาม, นาม, กริยา, คุณศัพท์	ประเภทของคำที่บอกให้รู้ว่าคำที่อ้างถึงเป็นเพศชาย เพศหญิง ไม่มีเพศ เป็นสิ่งมีชีวิตหรือไม่มีชีวิต ฯลฯ
การก (case)	สรรพนาม, กริยา, คุณศัพท์	แสดงถึง grammatical function ของคำในวลี อนุประโยค หรือประโยค ตัวอย่างเช่น ทำหน้าที่เป็นประธาน, (I ใน "I kicked the ball"), กรรม (me ใน "John kicked me"), แสดงความเป็นเจ้าของ ("My ball") เป็นต้น ตัวอย่าง case ที่พบบ่อยในภาษาอังกฤษ nominative case (e.g. subjective pronouns as I, he, she, we) accusative/dative case (e.g. objective pronouns as me, him, her, us) genitive case (e.g. such possessive pronouns as my/mine, his, her(s), our(s))

ข้อสนเทศของคำ (cont.)

ข้อสนเทศ	ใช้กับ	แสดง
กาล (tense)	กริยา	บอกถึงช่วงเวลาของการกระทำ โดยแบ่งออกเป็นอดีตกาล ปัจจุบันกาล และอนาคตกาล
การณ์ลักษณะ (aspect)	กริยา	หมายรวมถึงความสมบูรณ์ ความเป็นปกติวิสัย การดำเนินต่อไป เช่น imperfect (simple, progressive), perfect
มาลา (mood)	กริยา	สื่อความหมายในแง่ของข้อเท็จจริง ความเป็นไปได้ ความไม่แน่นอน เป็นต้น
วาจก (voice)	กริยา	ความสัมพันธ์ระหว่างประธานหรือสิ่งที่เกี่ยวข้องกับกริยาเพื่อแสดงการกระทำ แบ่งเป็น active, passive, middle, causative

Word and Language



Types of Language

- Analytic language is a language that primarily conveys relationships between words in sentences by way of helper words (particles, prepositions, etc.) and word order, as opposed to utilizing inflections (changing the form of a word to convey its role in the sentence)
- Synthetic language uses inflection or agglutination to express syntactic relationships within a sentence.
  - Inflection is the addition of morphemes to a root word that assigns grammatical property to that word
  - Agglutination is the combination of two or more morphemes into one word.

Inflectional and Derivational

- Inflection
  - The modification of a word to express different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case.
    - Conjugation is the inflection of verbs. For example, “go/goes/went/gone/going”
    - Declension is the inflection of nouns, adjectives and pronouns. For example, “he/his/him”, “book, books”
- Derivation
  - Used to form new words
    - *happiness* and *unhappy* from *happy*
    - *determination* from *determine*
    - *rewrite from write*
    - *Incorrect from correct*
  - Different from inflection in that, inflection uses another kind of affix in order to form variants of the same word

Derivational morpheme changes lexical categories of words, but inflectional morpheme does not.

WORD STATISTICS

How many words?

- **Types**: number of distinct words in a **corpus**
- **Vocabulary size**: number of distinct words
- **Token**: number of running words
- Example:
  - "**Natural Language processing** is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages."
  - How many word-type and token??

- **What is corpus??**
  - A *corpus* is a large collection of authentic (written or spoken) texts that have been gathered in electronic form.

Word counts

- Suppose that our text is represented as a list of words
- What are the most common words in the text?
- Example:

Book:"Tom Sawyer"

Word	Freq	Use
the	3332	Determiner (article)
and	2972	Conjunction
a	1775	Determiner
to	1725	Preposition, verbal infinitive marker
of	1440	Preposition
was	1161	Auxiliary verb
it	1027	Pronoun
in	906	Preposition
that	877	Complementizer, demonstrative
he	877	Pronoun
I	783	Pronoun
his	772	Possessive pronoun
you	686	Pronoun
Tom	679	Proper noun
with	642	preposition

Word counts (cont.)

- From the previous list
  - Dominated by the little words of English which have important grammatical roles
  - They are usually referred to as “function words”
    - Function words are, such as determiners, prepositions, complementizers
  - Tom!!! → proper noun
    - Frequency clearly reflects the text that we chose -> “Tom Sawyer”
- Frequency depends on the corpus or sample used.
  - What will happen if we change from ‘Entertainment news’ to ‘Political news’
  - Any difference? If we use political news from April-May 2010 and November-December 2008

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

Homework

- Group homework (around 4 students per group)
  - Link for downloading corpus: https://tinyurl.com/y2a7mysj
  - What you have to do
    - Count frequency of each word and sort them from highest counts to lowest counts (convert words to lowercase before counting)
    - Create frequency of frequency table and sort them from highest to lowest

Freq.	word
20	can, that, this
10	I, said, go, my
5	time, type, food, buy, kind
4	building, mother, find, miss, lost
2	friend, cat, dog, help, bus, car, sorry
1	school, hear, teach, present, correct, loud, traffic, light

Example

Freq.	count
1	8
2	7
4	5
5	5
10	4
20	3