Corpus-Based Work

สุธี สุดประเสริฐ

จากหนังสือ Foundations of Statistical Natural Langauge Processing บทที่ 4

Topics

- Getting Set Up
  - Computers
  - Corpora
  - Software
- Looking at Text
- Marked-up Data

Getting Set Up

- Computers
  - ข้อมูลที่นำมาวิเคราะห์ค่อนข้างมีขนาดใหญ่ จึงจำเป็นต้องใช้คอมพิวเตอร์ที่มีประสิทธิภาพการคำนวนสูง
  - ในสมัยก่อน Brown Corpus (1960s) ใช้เวลาในการเรียงลำดับคำถึง 17 ชั่วโมง
    - หน่วยความจำขนาด 40 KB อ่านข้อมูลจากเทป
  - งานส่วนใหญ่มักจะเป็นการเข้าไปนับคลังประโยคขนาดใหญ่
    - เนื้อที่ในการเก็บข้อมูลและหน่วยความจำขนาดใหญ่

Getting Set Up

- Corpora

Linguistic Data Consortium (LDC)	http://www.ldc.upenn.edu
European Language Resources Associatin (ELRA)	http://elra.info
International Computer Archive of Modern English (ICAME)	http://icame.uib.no
Oxford Text Archive (OTA)	http://ota.ahds.ac.uk
Child Language Data Exchange System (CHILDES)	http://childes.psy.cmu.edu

- รายชื่อองค์กรที่จัดเตรียมคลังเอกสารเพื่อใช้สำหรับงานด้านภาษาศาสตร์โดยเฉพาะ
  - ถ้าต้องการใช้จำเป็นต้องเสียเงิน

Getting Set Up

- Corpora
  - มีคลังประโยคจำนวนมากบนเว็บ ที่ไม่จำเป็นต้องเสียเงินนำการใช้ เช่น
    - Brown corpus
    - Lancaster corpus
    - CSLU Speech corpus
  - ส่วนใหญ่จะไม่มีการกำกับข้อมูลทางภาษาศาสตร์ลงไป (ข้อมูลดิบ)
  - อย่างไรก็ตาม การทำงานบนข้อมูลดิบเป็นความท้าทายอย่างหนึ่ง

Getting Set Up

- Corpora
  - การเลือกใช้คลังประโยคต้องคำนึงถึงความสมเหตุสมผลและผลลัพธ์ที่เราต้องการ
    - คลังประโยคแต่ละอัน ถูกสร้างขึ้นมาด้วยหลักการที่แตกต่างกัน
      - Brown corpus: ภาษาเขียนในปี 1961
    - การจัดกลุ่มเอกสารจำเป็นต้องเปลี่ยนข้อมูลที่นำมาเรียนรู้อยู่เรื่อยๆ
  - คลังประโยคต้องสามารถแทนกลุ่มตัวอย่าง (*representative sample*) ของข้อมูลที่เราสนใจได้
  - นอกจากนั้นคลังประโยคต้องมีความสมดุล (balanced corpus)

Getting Set Up

- Software
  - Text editors
  - Regular expression
  - Programming languages
  - Programming techniques
    - Coding words
    - Collecting count data

Getting Set Up

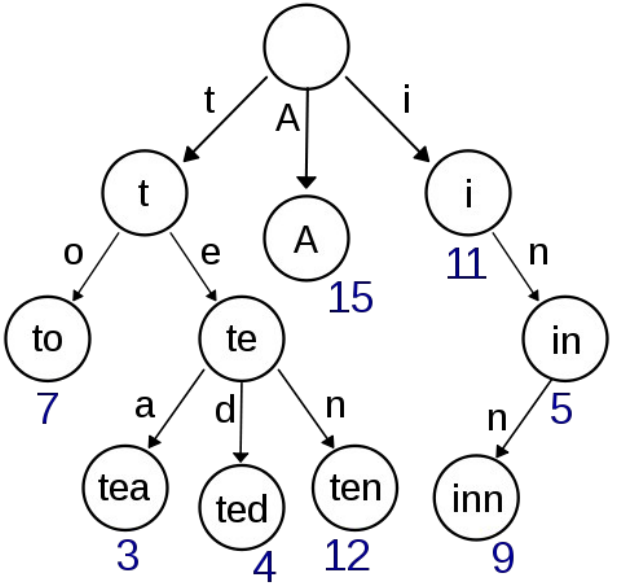
- Text editors
  - EditPlus, Notepad++, Vim, Emacs
- Regular expressions
  - ปกติจะมีอยู่ใน text editors (จะพูดอีกทีในบทต่อไป)
  - โปรแกรมภาษาส่วนใหญ่มี libraries ให้เรียกให้ได้
  - สามารถอธิบาย regular languages
  - มีความสามารถเท่ากับ finite state automata

Getting Set Up

- Programming languages
  - C/C++ : ใช้ในกรณีที่ต้องการประสิทธิภาพในการประมวล หรือ ต้องจัดการกับข้อมูลเป็นจำนวนมากๆ
  - Perl : เน้นงานด้านการจัดการกับสตริง
  - Python : คล้าย Perl แต่การจัดการกับสตริงจะด้อยกว่า แต่การเขียนโปรแกรมจะเป็นระบบและง่ายกว่า และมี modules ศึกษาการพัฒนาโปรแกรมด้าน NLP จำนวนมาก (เช่น NLTK)
  - Prolog : เน้นงานด้านการให้เหตุผลและระบบผู้เชี่ยวชาญ

Getting Set Up

- Programming techniques
  - Coding words
    - โดยปกติเวลาทำงานจริงเราจะไม่เก็บคำไว้ตรงๆ แต่จะใช้การแปลงคำเป็นตัวเลขก่อน โดยใช้ hash table
    - วิธีการอื่นๆ ที่สามารถใช้ได้ เช่น trie หรือ prefix tree



http://en.wikipedia.org/wiki/Trie

Getting Set Up

- Programming techniques
  - Collecting count data
    - ไม่จำเป็นต้องนับจากข้อมูลทั้งหมด
    - สร้างตัวแทนส่วนที่ต้องการนับก่อน จากนั้นจึงเรียงลำดับใหม่แล้วจึงค่อยนับ
    - เครื่องมือช่วย : NLTK, CMU-Cambridge Statistical Language Modeling toolkit (Unix)

Looking at Text

- เอกสารที่นำมาใช้อาจจะเป็นข้อมูลดิบหรือการกำกับข้อมูล (marked up) บางส่วนเอาไว้
  - Markup คือส่วนที่ไม่ใช่ข้อมูลของเอกสาร แต่ใส่ไว้เพื่อแสดงข้อมูลบางอย่าง
- ภาษาธรรมชาติมีคุณสมบัติหลายอย่างที่ทำให้ การนำเอกสารมาประมวลผลทำได้ยาก

Low-level formatting issues

- Junk formatting/content
  - ข้อมูลต่างๆ ที่ไม่ต้องการและจำเป็นต้องลบทิ้งก่อนจะนำเอกสารมาประมวลผล
    - รูปภาพ ตาราง ตัวแบ่งหน้า หัว/ท้ายเอกสาร ตัวอักขระพิเศษ ฯลฯ
- Uppercase and lowercase
  - เฉพาะภาษาอังกฤษ

Tokenization: What is a word?

- โดยปกติ ขั้นตอนแรกสุดในการประมวลผลเอกสารคือการแบ่งข้อความเป็นส่วนๆ ซึ่งเรียกว่า *tokens*
  - คำ ตัวเลข สัญลักษณ์
- ส่วนใหญ่จะเก็บข้อมูลการกำหนดขอบเขตของประโยคเอาไว้ รวมทั้งสัญลักษณ์ต่างๆ เช่น , และ -
  - ข้อมูลเหล่านี้อาจนำไปใช้ในการวิเคราะห์ในขั้นตอนต่อไป

Tokenization: What is a word?

- Periods (.)
  - ในภาษาอังกฤษ . ใช้ในบอกขอบเขตของประโยคและการเขียนตัวย่อ ในบางกรณีอาจการความสับสนได้
- Single apostrophes (')
  - I'll หรือ isn't ควรเป็นหนึ่งคำหรือสองคำ
  - dog's ไม่ได้หมายถึง dog is เสมอไป
- Hyphenation (-)
  - text-based ควรเป็นหนึ่งคำหรือสองคำ
  - บางเอกสารใช้ในการแบ่งคำเพื่อขึ้นบรรทัดใหม่

Tokenization: What is a word?

- Word segmentation
  - ในบางภาษาไม่มีตัวบ่งบอกขอบเขตของคำ
    - ไทย จีน ญี่ปุ่น เกาหลี ฯลฯ
  - ในบางภาษาคำนามผสมจะเขียนติดกัน
    - เยอรมัน : Lebensversicherungsgesellschaftsangestellter (life insurance company employee)
  - ในภาษาอังกฤษคำนามผสมบางคำเขียนติดกัน
    - data base = database
    - hard disk = harddisk

Tokenization: What is a word?

- ในภาษาอังกฤษช่องว่างไม่ได้ใช้สำหรับการแบ่งคำเสมอไป
  - Phrasal verbs
    - make up, work out, give up, pick up, …
  - Multipart names
    - New York, San Francisco, …
  - Fixed phrases
    - in spite of, in order to, because of, ...

Tokenization: What is a word?

- ความหลากหลายในการเขียนแสดงข้อมูลที่มีความหมายเดียวกัน
  - หมายเลขโทรศัพท์
    - 08-456-6666, (+66) 8456-6666, 084566666, ...
  - วันที่
    - 12/12/53, 12/12/2553, 12 ธค. 2553, …
  - เวลา
    - 12.00 น., 12 นาฬิกา, เที่ยงวัน

Morphology

- จะพูดละเอียดอีกทีในบทต่อไป
  - sit, sits, sat จะเก็บแยกหรือเก็บรวมกัน

Sentences

- What is a sentence?
  - สำหรับภาษาอังกฤษ อาจบอกว่า สตริงที่ลงท้ายด้วย . ? หรือ !
  - ประโยคในภาษาอังกฤษ 90% จะลงท้ายด้วย .
  - สำหรับภาษาไทย จะใช้ช่องว่างในการแบ่งประโยค ซึ่งจะมีความกำกวมมากกว่า
    - ช่องว่างถูกใช้ในหลายวัตถุประสงค์
      - แบ่งกลุ่มสำหรับตัวอย่าง
      - หยุดข้อความเพื่อให้อ่านง่ายขึ้น
    - หรืออาจบอกว่าในภาษาไทย ไม่มีกฎตายตัวในการใช้ช่องว่าง

Sentences

“You remind me,” she remarked, “of your dad.”

ในการประชุมคณะกรรมการประจำคณะวิทยาศาสตร์ ครั้งที่ 18/2553 เมื่อวันพุธที่ 13 ตุลาคม 2553 ได้มีการปรับวงเงินใหม่เป็นครั้งที่ 4

ข้อความเหล่านี้มีกี่ประโยค?

Sentences

Lenght	Number	Percent	Cum. %
1-5	1317	3.13	3.13
6-10	3215	7.64	10.77
11-15	5906	14.03	24.80
16-20	7206	17.12	41.92
21-25	7350	17.46	59.38
26-30	6281	14.92	74.30
31-35	4740	11.26	85.56
36-40	2826	6.71	92.26
41-45	1606	3.82	96.10
46-50	858	2.04	98.14
51-100	780	1.85	99.99
101+	6	0.01	100.00

สถิติข้อมูลความยาวของประโยคอังกฤษโดยรวบรวมจากข่าวหนังสือพิมพ์

Marked-up Data

- หากเอกสารที่เรานำมาประมวลมีการกำกับ "ข้อมูลทางภาษาศาสตร์" เอาไว้จะทำให้การประมวลผลง่ายขึ้น
  - พื้นฐาน: ขอบเขตคำ ประโยค ย่อหน้า
  - ขั้นสูง: วาจก (POS), ความหมาย, โครงสร้างไวยากรณ์
- การกำกับข้อมูล สามารถให้คนกำกับ หรือ กำกับด้วยเครื่องโดยอัตโนมัติ หรือ ใช้ทั้งสองวิธีผสมกัน

Markup schemes

- SGML (Standard Generalized Markup Language)
  - HTML, XML

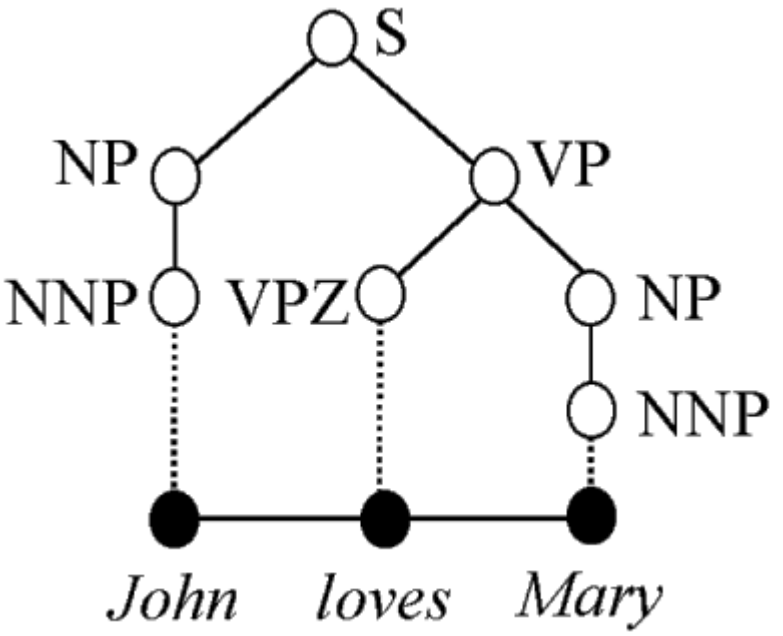
```
<p>
 <s>And then he left.</s>
 <s>He did not say another word.</s>
</p>

<utt speak="Fred" date="10-Feb-1998">
   That is an ugly couch.
</utt>
```

Markup schemes

- Lisp-like bracketing

```
(S  (NP  (NNP John))
    (VP  (VPZ loves)
         (NP  (NNP Mary)))
    (.  .))
```



Grammatical tagging

- Tag sets

Sentence	CLAWS c5	Brown	Penn TB
she	PNP	PPS	PRP
was	VBD	BEDZ	VBD
told	VVN	VBN	VBN
that	CJT	CS	IN
the	AT0	AT	DT
journey	NN1	NN	NN
might	VM0	MD	MD
kill	VVI	VB	VB
her	PNP	PPO	PRP
.	PUN	.	.

Brown = 87 tags
Penn TB = 45 tags
CLAWS c5 = 62 tags

Grammatical tagging

Category	Examples	Claws c5	Brown	Penn TB
Adjective	happy, bad	AJ0	JJ	JJ
Adjective, ordinal number	sixth, 72nd, last	ORD	OD	JJ
Adjective, comparative	happier, worse	AJC	JJR	JJR
Adjective, superlative	happiest, worst	AJS	JJT	JJS
Adjective, superlative, semantically	chief, top	AJ0	JJS	JJ

The design of a tag set

- Features for guiding the design of a tag set
  - The feature of classification
    - บอกถึงข้อมูลที่มีประโยชน์เกี่ยวกับชนิดของคำในเชิงไวยากรณ์
  - The predictive features
    - คาดเดาพฤติกรรมของคำอื่นๆ ในบริบทได้
- แต่ในความเป็นจริง สองคุณลักษณะนี้มักจะขัดแย้งกันเอง เช่น -ing ควรเป็น verb หรือ noun
  - Verb : แบ่งหมวดหมู่ได้ชัดเจน
  - Noun : บริบทใช้เหมือนคำนาม