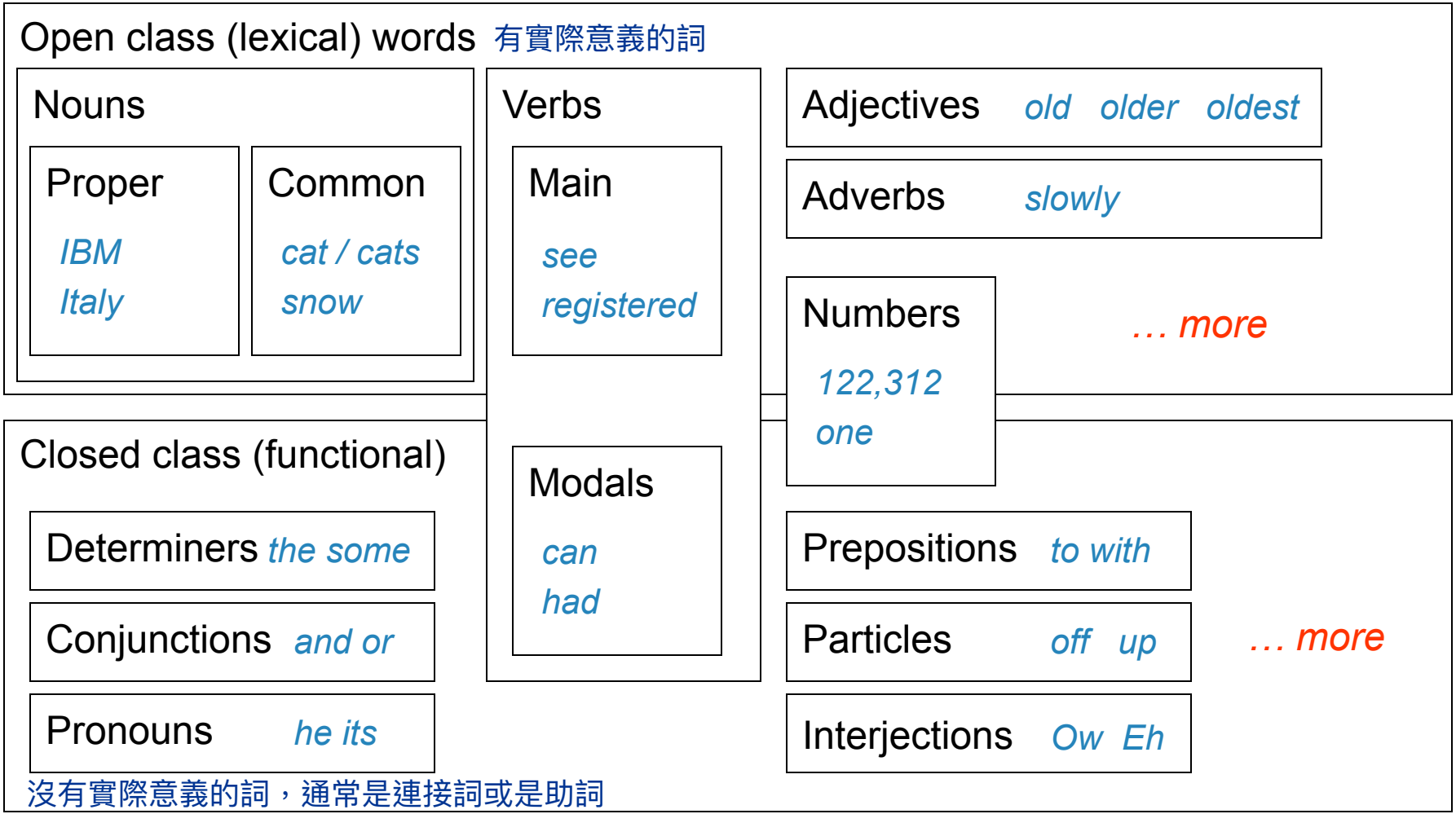
Part-of-speech tagging

A simple but useful form of linguistic analysis

Christopher Manning

Parts of Speech 詞性辨識

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
  - a.k.a lexical categories, word classes, "tags", POS
- It comes from Dionysius Thrax of Alexandria (c. 100 BCE) the idea that is still with us that there are 8 parts of speech
  - But actually his 8 aren't exactly the ones we are taught today
    - Thrax: noun, verb, article, adverb, preposition, conjunction, participle, pronoun
    - School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection


Open class (lexical) words 有實際意義的詞
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
Modals
can
had
Prepositions to with
Particles off up
Interjections Ow Eh
… more
沒有實際意義的詞，通常是連接詞或是助詞

Open vs. Closed classes

- Open vs. Closed classes
  - Closed:
    - determiners: ***a, an, the***
    - pronouns: ***she, he, I***
    - prepositions: ***on, under, over, near, by, …***
    - Why "closed"?
  - Open:
    - Nouns, Verbs, Adjectives, Adverbs.

POS Tagging

同一個字可能會有很多詞性，所以要看前後文才能決定

- Words often have more than one POS: *back*
  - The *back* door = JJ ADJ
  - On my *back* = NN N
  - Win the voters *back* = RB ADV
  - Promised to *back* the bill = VB V
- The POS tagging problem is to determine the POS tag for a particular instance of a word.

POS Tagging

一個字會有很多種詞性，POS Tagging的用途是找出正確的詞性組合

- Input: Plays well with others
- Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS
- Output: Plays/VBZ well/RB with/IN others/NNS
- Uses:
  - Text-to-speech (how do we pronounce “lead”?)
  - Can write regexps like (Det) Adj* N+ over the output for phrases, etc.
  - As input to or to speed up a full parser
  - If you know the tag, you can back off to it in other tasks

Penn Treebank POS tags

POS tagging performance

- How many tags are correct? (Tag accuracy)
  - About 97% currently
  - But baseline is already 90%
    - Baseline is performance of stupidest possible method
      - Tag every word with its most frequent tag
      - Tag unknown words as nouns
  - Partly easy because
    - Many words are unambiguous
    - You get points for them (*the*, *a*, etc.) and for punctuation marks!

Deciding on the correct part of speech can be difficult even for people

困難的POS Tagging句子例子

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD

How difficult is POS tagging?

- About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., *that*
  - I know *that* he is honest = IN
  - Yes, *that* play was nice = DT
  - You can't go *that* far = RB
- 40% of the word tokens are ambiguous

Part-of-speech tagging

A simple but useful form of linguistic analysis

Christopher Manning

Part-of-speech tagging revisited

A simple but useful form of linguistic analysis

Christopher Manning

Sources of information

FEATURES

- What are the main sources of information for POS tagging?
  - Knowledge of neighboring words
    - Bill saw that man yesterday
    - NNP NN DT NN NN
    - VB VB(D) IN VB NN
  - Knowledge of word probabilities
    - *man* is rarely used as a verb….
- The latter proves the most useful, but the former also helps

這句話有2^4種詞性排列組合
但有很多其實是不合理的
例如Bill很少當作VB、that後面不會接VB

More and Better Features → Feature-based tagger

雖然很多字有多種詞性，但很多詞其實用字根就可以直接看出詞性

- Can do surprisingly well just looking at a word by itself:
  - Word: the: the → DT
  - Lowercased word: Importantly: importantly → RB
  - Prefixes: unfathomable: un- → JJ
  - Suffixes: Importantly: -ly → RB
  - Capitalization: Meridian: CAP → NNP
  - Word shapes: 35-year: d-x → JJ

中間有一個dash的就是形容詞

- Then build a maxent (or whatever) model to predict tag
  - Maxent P(t|w): 93.7% overall / 82.6% unknown

Overview: POS Tagging Accuracies

- Rough accuracies:

	overall / unknown words
• Most freq tag:	~90% / ~50%
• Trigram HMM:	~95% / ~55%
• Maxent P(t|w): 最直覺的方法	93.7% / 82.6%
• TnT (HMM++):	96.2% / 86.0%
• MEMM tagger:	96.9% / 86.9%
• Bidirectional dependencies:	97.2% / 90.0%
• Upper bound:	~98% (human agreement)

Most errors on unknown words

How to improve supervised results?

- Build better features!

		RB					
PRP	VBD	IN	RB	IN	PRP	VBD	.
They	left	as	soon	as	he	arrived	.

答案應該是RB
但卻誤標為IN
那要怎麼修正這個錯誤？
可以在標記時多看下一個字
as soon
as是拿來修飾soon(RB)
所以會是副詞(RB)而非(IN)

- We could fix this with a feature that looked at the next word

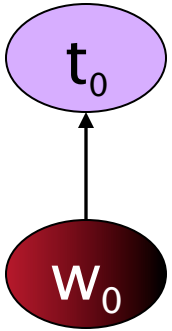
JJ					
NNP	NNS	VBD	VBN	.	
Intrinsic	flaws	remained	undetected	.	

對於沒看過的字Intrinsic
因為首字母大寫，先標記成名詞
後來發現intrinsic是形容詞
就再改成形容詞

- We could fix this by linking capitalized words to their lowercase versions

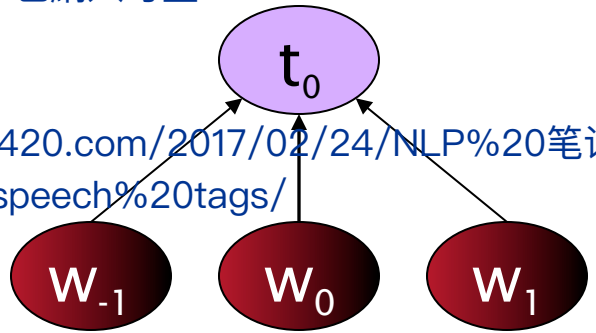
Tagging Without Sequence Information

Baseline



Three Words

如果把目標字的前後字也納入考量
效果會提升
參考：
http://www.shuang0420.com/2017/02/24/NLP%20笔记%20-%20Part%20of%20speech%20tags/



Model	Features	Token	Unknown	Sentence
Baseline	56,805	**93.69%**	82.61%	26.74%
3Words	239,767	**96.57%**	86.78%	48.27%

Using words only in a straight classifier works as well as a basic (HMM or discriminative) sequence model!!

Summary of POS Tagging

For tagging, the change from generative to discriminative model **does not by itself** result in great improvement

One profits from models for specifying dependence on **overlapping features of the observation** such as spelling, suffix analysis, etc.

An MEMM allows integration of rich features of the observations, but can suffer strongly from assuming independence from following observations; this effect can be relieved by adding dependence on following words

This additional power (of the MEMM ,CRF, Perceptron models) has been shown to result in improvements in accuracy

The **higher accuracy** of discriminative models comes at the price of **much slower training**

Part-of-speech tagging revisited

A simple but useful form of linguistic analysis

Christopher Manning