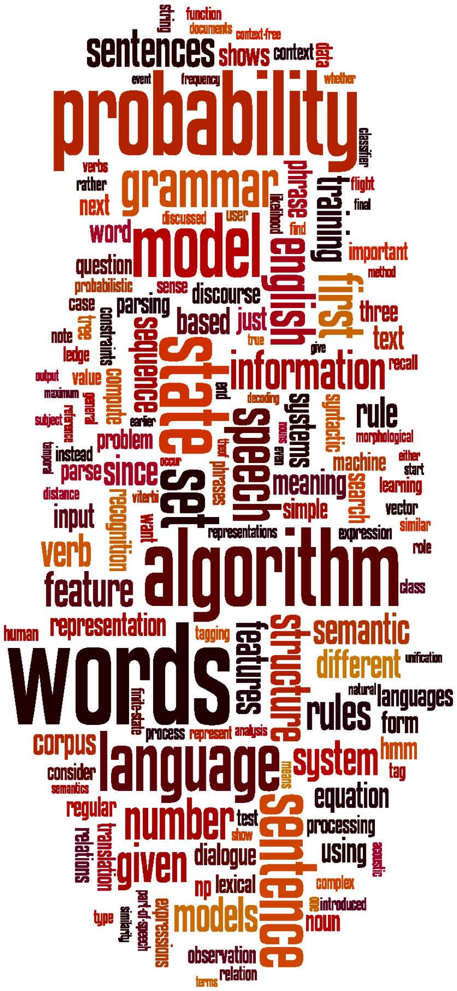


Information Extraction and Named Entity Recognition

Introducing the tasks:

Getting simple structured information out of text

Information Extraction

目標：把雜亂的文章擷取出重要的資訊並儲存為結構化的資料

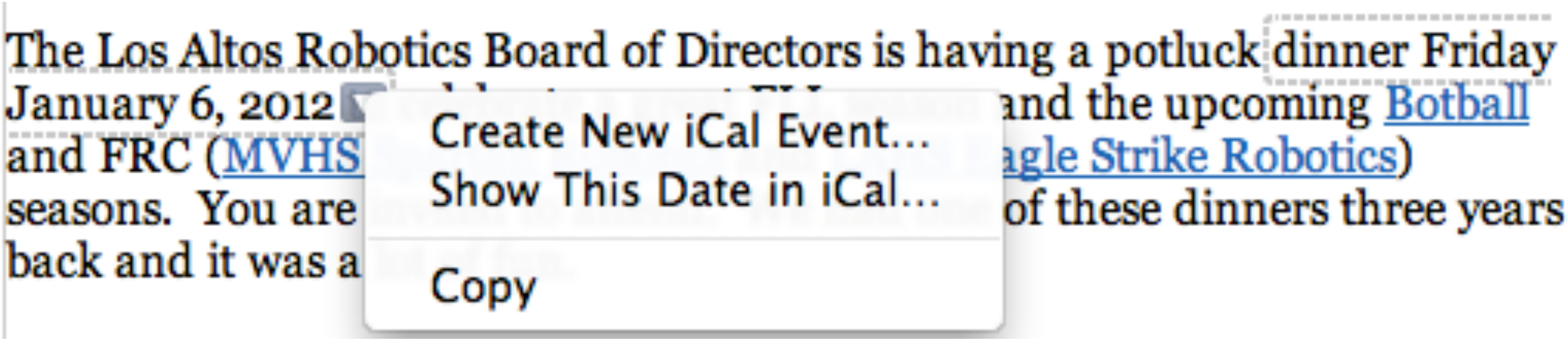
- Information extraction (IE) systems
  - Find and understand limited relevant parts of texts
  - Gather information from many pieces of text
  - Produce a structured representation of relevant information:
    - *relations* (in the database sense), a.k.a.,
    - a *knowledge base*
  - Goals:
    1. Organize information so that it is useful to people
    2. Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction (IE)

- IE systems extract clear, factual information
  - Roughly: *Who did what to whom when?*
- E.g.,
  - Gathering earnings, profits, board members, headquarters, etc. from company reports
    - The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.
    - headquarters(“BHP Biliton Limited”, “Melbourne, Australia”)
  - Learn drug-gene product interactions from medical research literature

Low-level information extraction

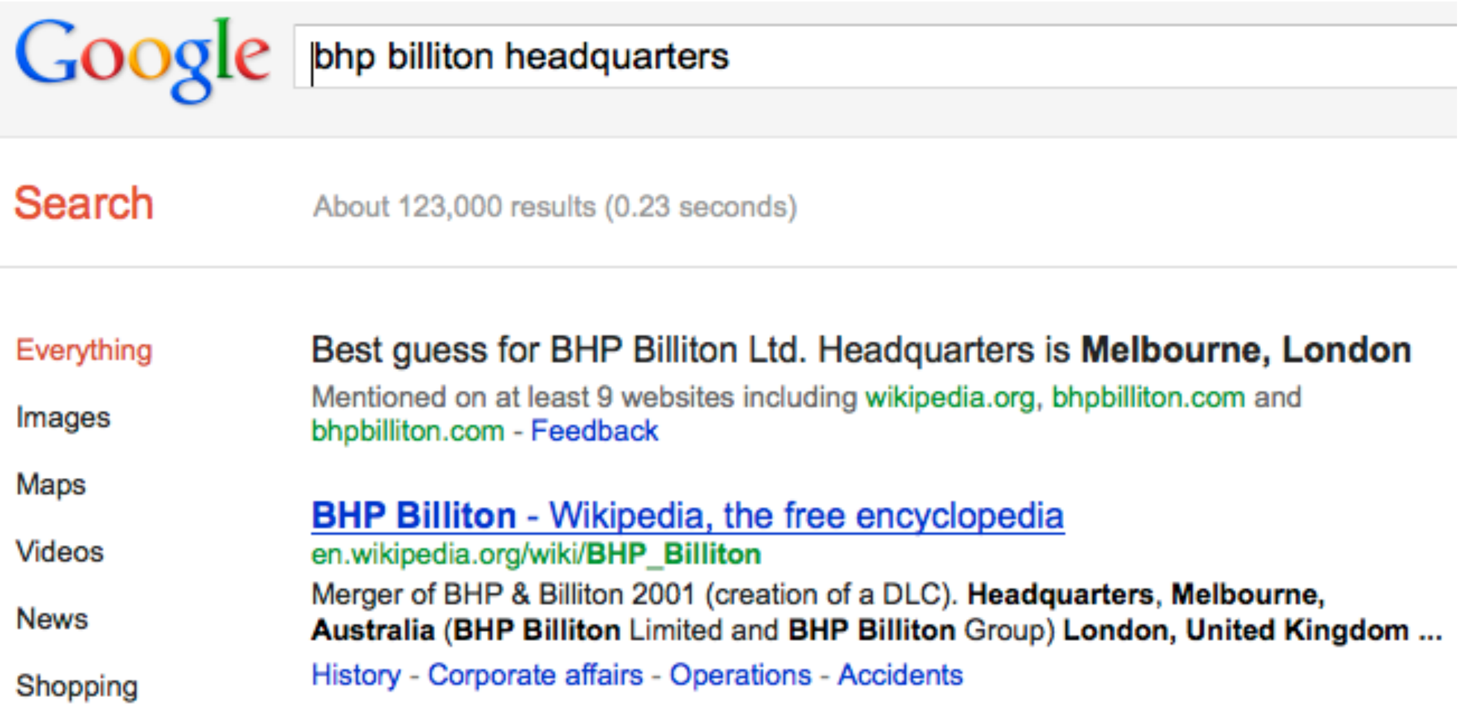
- Is now available – and I think popular – in applications like Apple or Google mail, and web indexing



The Los Altos Robotics Board of Directors is having a potluck dinner Friday January 6, 2012 and the upcoming Botball and FRC (MVHS agle Strike Robotics) seasons. You are of these dinners three years back and it was a

- Often seems to be based on regular expressions and name lists

Low-level information extraction



bhp billiton headquarters

About 123,000 results (0.23 seconds)

Best guess for BHP Billiton Ltd. Headquarters is **Melbourne, London**

Mentioned on at least 9 websites including wikipedia.org, bhpbilliton.com and bhpbilliton.com - Feedback

BHP Billiton - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/**BHP_Billiton**

Merger of BHP & Billiton 2001 (creation of a DLC). **Headquarters**, **Melbourne, Australia** (**BHP Billiton** Limited and **BHP Billiton** Group) **London, United Kingdom** ...

History - Corporate affairs - Operations - Accidents

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  標記出命名實體，並分類出該命名實體是人名、地名、組織名等等

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:
  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

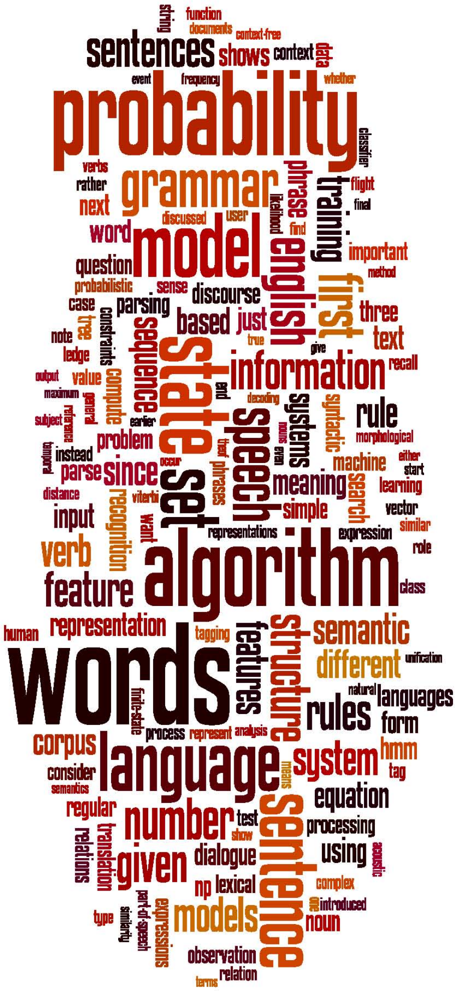
Named Entity Recognition (NER)

- The uses:
  - Named entities can be indexed, linked off, etc.
  - Sentiment can be attributed to companies or products
  - A lot of IE relations are associations between named entities
  - For question answering, answers are often named entities.
- Concretely:
  - Many web pages tag various entities, with links to bio or topic pages, etc.
    - Reuters' OpenCalais, Evri, AlchemyAPI, Yahoo's Term Extraction, …
  - Apple/Google/Microsoft/… smart recognizers for document content

Information Extraction and Named Entity Recognition

Introducing the tasks:

Getting simple structured information out of text



Evaluation of Named Entity Recognition

The extension of Precision, Recall, and the F measure to sequences

The Named Entity Recognition Task

Task: Predict entities in a text

	predict entity
Foreign	ORG
Ministry	ORG
spokesman	O
Shen	PER
Guofang	PER
told	O
Reuters	ORG
:	:

} Standard evaluation is per entity, *not* per token

有可能兩個單字組成一個Entity
但傳統的衡量方法是判斷每一個token
所以可能會高估NER的成果

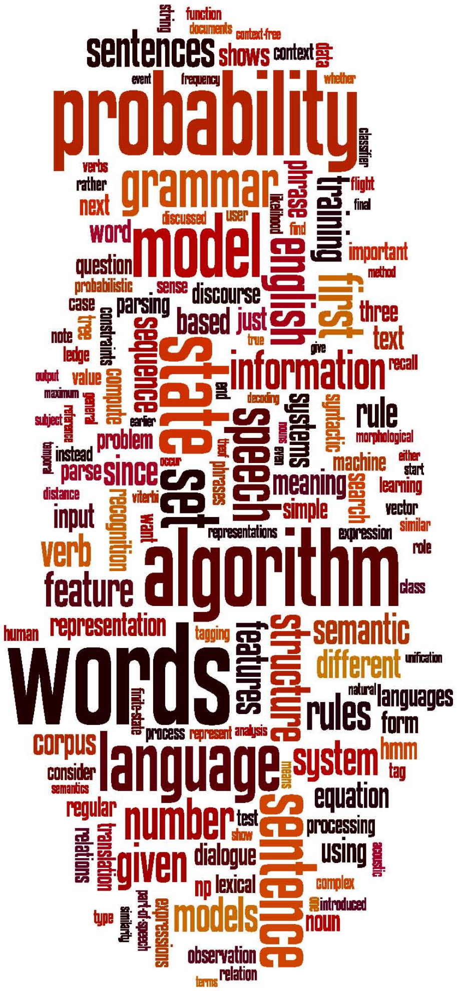
Precision/Recall/F1 for IE/NER

- Recall and precision are straightforward for tasks like IR and text categorization, where there is only one grain size (documents)
- The measure behaves a bit funnily for IE/NER when there are *boundary errors* (which are *common*):
  - First Bank of Chicago announced earnings …
- This counts as both a fp and a fn
- Selecting *nothing* would have been better
- Some other metrics (e.g., MUC scorer) give partial credit (according to complex rules) 計算部分得分的演算法

正解：First Bank of Chicago
標記：Bank of Chicago
這是一種boundary error
left boundary錯誤
right boundary正確

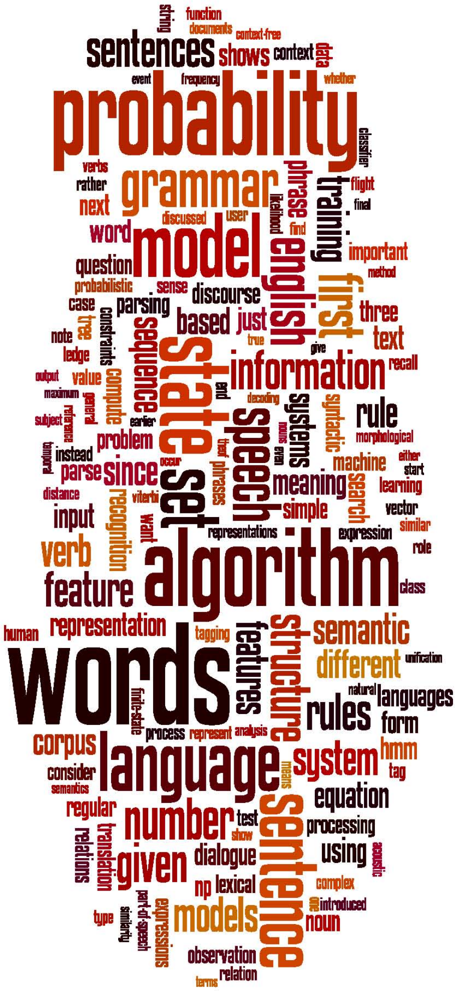
什麼都沒標記，只會有一個錯誤fn
會降低Recall，但不影響Precision
但如果標記錯了，會有兩個錯誤fp & fn
會同時降低Recall和Precision

tp(truth positive):選到正確的東西
fp(false positive):選到錯誤的東西
fn(false negative):沒選到正確的東西
tn(truth negative):沒選到錯誤的東西



Evaluation of Named Entity Recognition

The extension of Precision, Recall, and the F measure to sequences



Sequence Models for Named Entity Recognition

The ML sequence model approach to NER

Training 選定資料集–>標記–>設計特徵–>訓練

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
‘s	O	O
new	O	O
painting	O	O

IO encoding
只有區分Entity和非Entity
這樣的問題是如果有長度大於一的Entity
就可能會沒辦法辨認出來
所以有了IOB encoding

IOB encoding
B: Beginning，代表Entity的開始
I: Inside，代表Entity的中間

例如
Mengqiu Huang是一個人
但在IO encoding會被視為兩個人
在IOB encoding就會被視為一個人

Features for sequence labeling

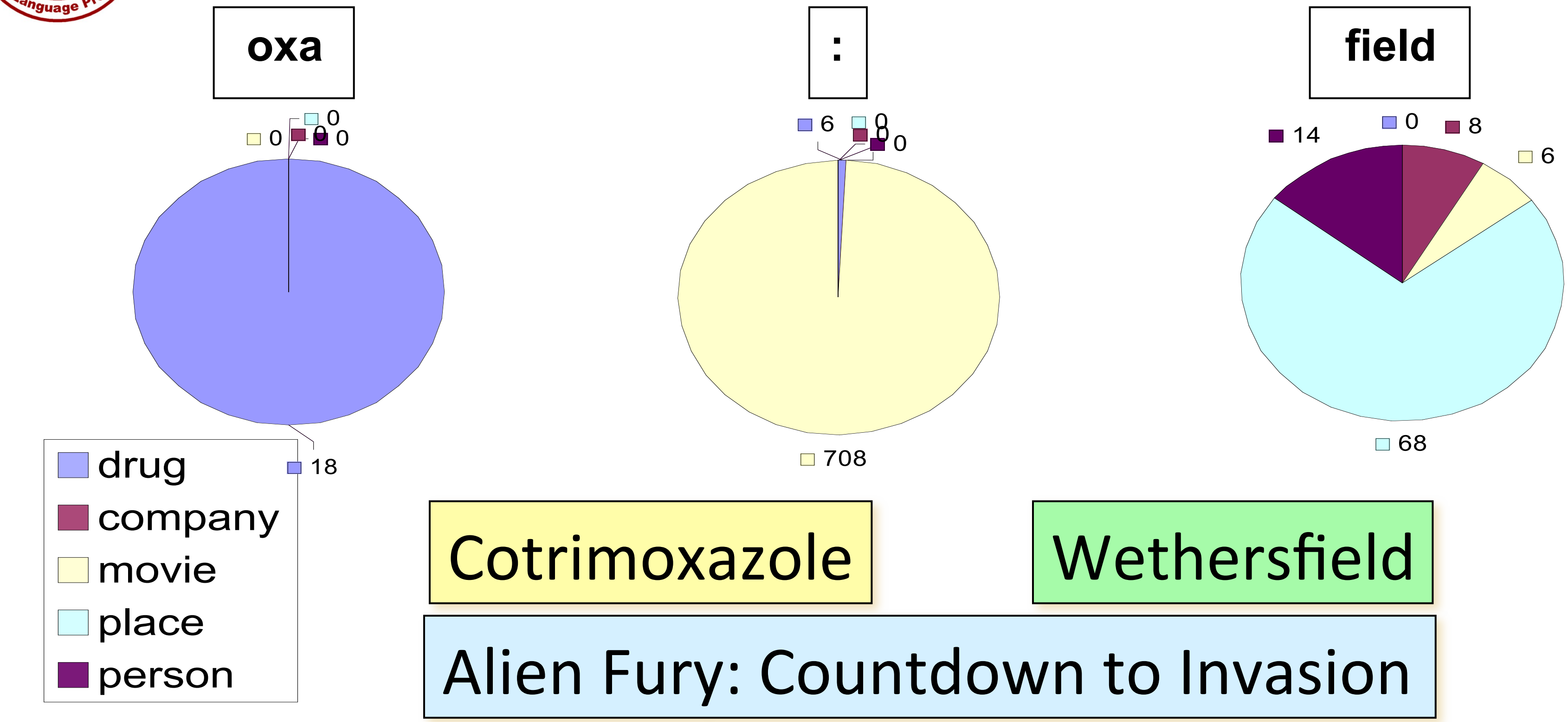
- Words
  - Current word (essentially like a learned dictionary)
  - Previous/next word (context)

suffix
prefix
特徵可以包含Entity的前後文

- Other kinds of inferred linguistic classification
  - Part-of-speech tags 詞性標記
- Label context
  - Previous (and perhaps next) label

用字根來判斷一個字的種類
例如：看到oxa，代表該字可能代表某種藥物
看到field，代表該字可能是個地點

Features: Word substrings



Cotrimoxazole

Wethersfield

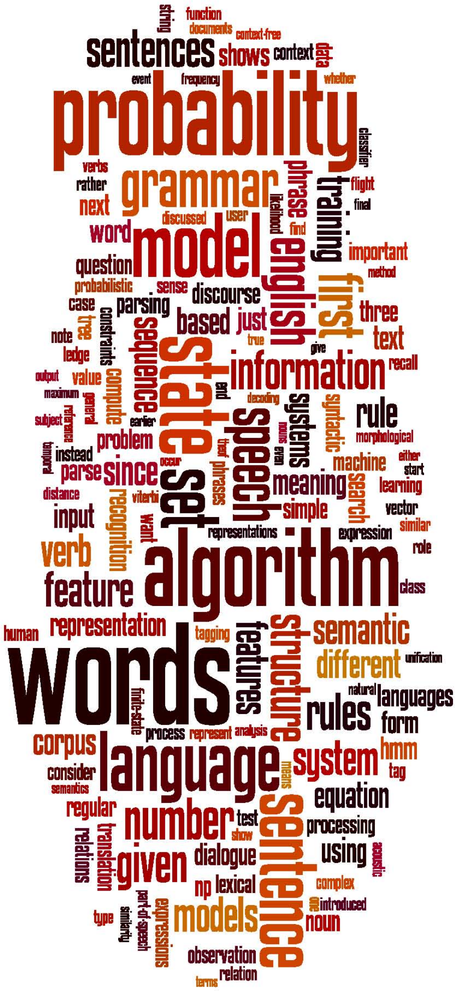
Alien Fury: Countdown to Invasion

Features: Word shapes

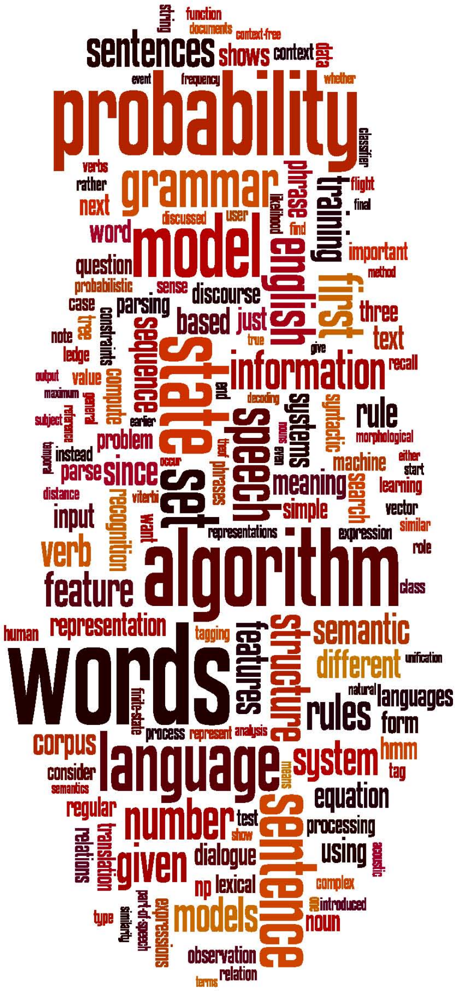
- Word Shapes
  - Map words to simplified representation that encodes attributes such as length, capitalization, numerals, Greek letters, internal punctuation, etc.

Varicella-zoster	Xx-xxx
mRNA	xXXX
CPA1	XXXd

用類似正規表示式的方式
來辨認出一些具有特殊格式的專有名詞等Entity



Sequence Models for Named Entity Recognition



Maximum entropy sequence models

Maximum entropy Markov models (MEMMs) or Conditional Markov models

Sequence problems

Sequence Labeling 應用

- Many problems in NLP have data which is a sequence of characters, words, phrases, lines, or sentences …
- We can think of our task as one of labeling each item

VBG	NN	IN	DT	NN	IN	NN
Chasing	opportunity	in	an	age	of	upheaval

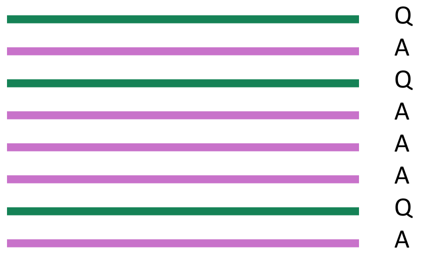
POS tagging 詞性標記

PERS	O	O	O	ORG	ORG
Murdoch	discusses	future	of	News	Corp.

Named entity recognition 命名實體辨識

B	B	I	I	B	I	B	I	B	B
而	相	对	于	这	些	品	牌	的	价

Word segmentation 斷詞



Text segmentation

MEMM inference in systems

- For a Conditional Markov Model (CMM) a.k.a. a Maximum Entropy Markov Model (MEMM), the classifier makes a single decision at a time, conditioned on evidence from observations and previous decisions
- A larger space of sequences is usually explored via search

Local Context

Decision Point

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

用MEMM去做Sequence Labeling
那要怎麼判斷每個Token的tag呢？
可以設計Feature，包含前n個字和後n個字的內容與tag
或是該token有沒有包含英文字、數字等

Example: POS Tagging

- Scoring individual labeling decisions is no more complex than standard classification decisions
  - We have some assumed labels to use for prior positions
  - We use features of those and the observed data (which can include current, previous, and next words) to predict the current label

Local Context

Decision Point

-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

Example: POS Tagging

- POS tagging Features can include:
  - Current, previous, next words in isolation or together.
  - Previous one, two, three tags.
  - Word-internal features: word types, suffixes, dashes, etc.

Local Context

Decision Point

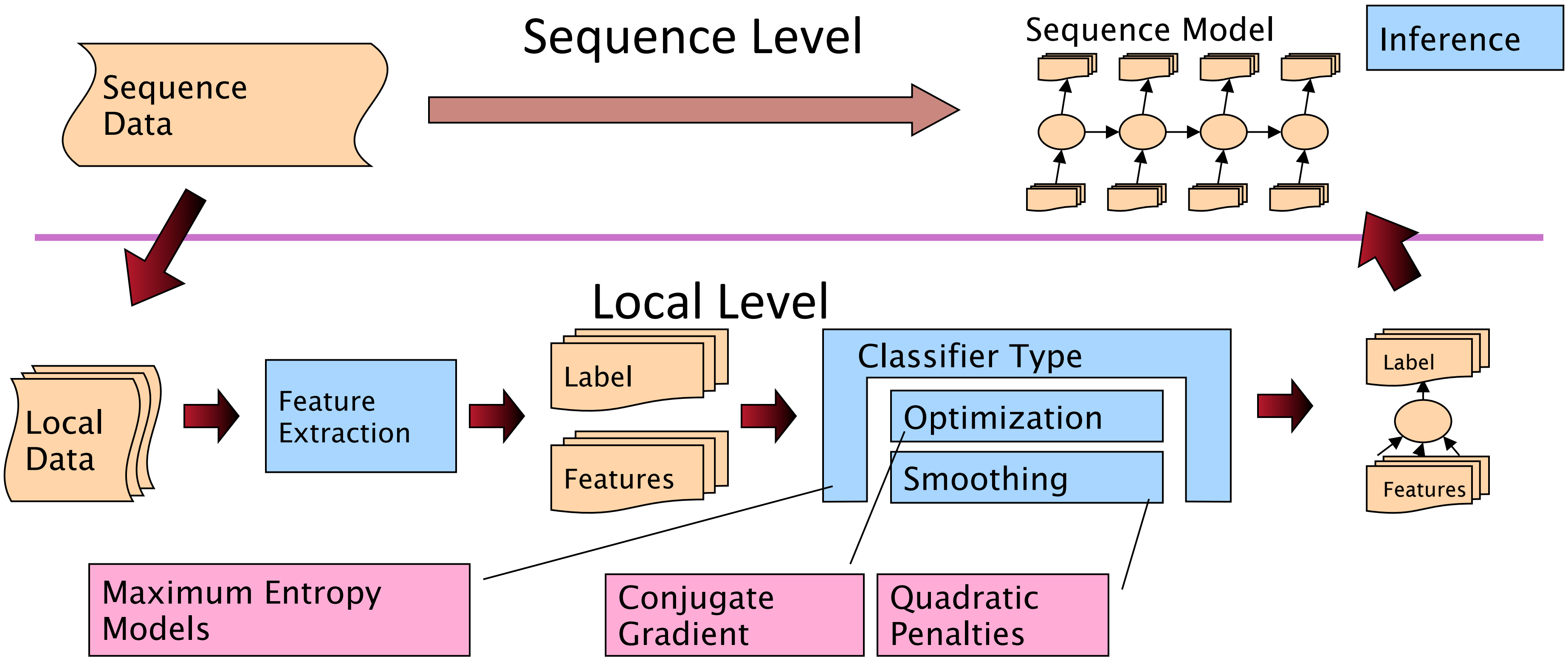
-3	-2	-1	0	+1
DT	NNP	VBD	???	???
The	Dow	fell	22.6	%

Features

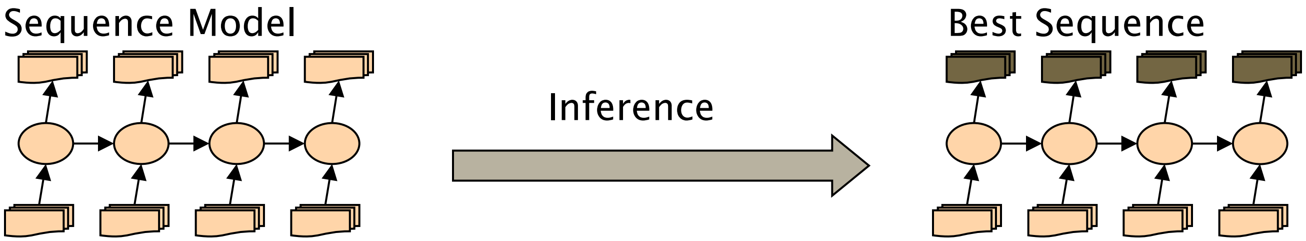
W_0	22.6
W_{+1}	%
W_{-1}	fell
T_{-1}	VBD
T_{-1}-T_{-2}	NNP-VBD
hasDigit?	true
…	…

(Ratnaparkhi 1996; Toutanova et al. 2003, etc.)

Inference in Systems



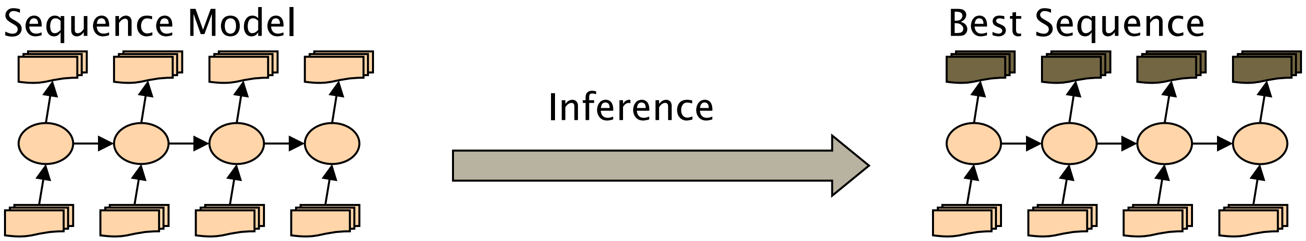
Greedy Inference



- Greedy inference:
  - We just start at the left, and use our classifier at each position to assign a label
  - The classifier can depend on previous labeling decisions as well as observed data
- Advantages:
  - Fast, no extra memory requirements
  - Very easy to implement
  - With rich features including observations to the right, it may perform quite well
- Disadvantage:
  - Greedy. We make commit errors we cannot recover from

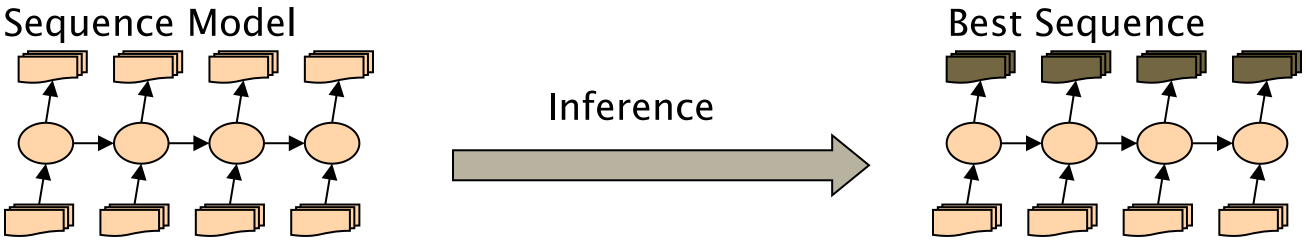
Christopher Manning

Beam Inference



- Beam inference:
  - At each position keep the top k complete sequences.
  - Extend each sequence in each local way.
  - The extensions compete for the k slots at the next position.
- Advantages:
  - Fast; beam sizes of 3–5 are almost as good as exact inference in many cases.
  - Easy to implement (no dynamic programming required).
- Disadvantage:
  - Inexact: the globally best sequence can fall off the beam.

Viterbi Inference



- Viterbi inference:
  - Dynamic programming or memoization.
  - Requires small window of state influence (e.g., past two states are relevant).
- Advantage:
  - Exact: the global best sequence is returned.
- Disadvantage:
  - Harder to implement long-distance state-state interactions (but beam inference tends not to allow long-distance resurrection of sequences anyway).

http://cpmarkchang.logdown.com/posts/192522-natural-language-processing-viterbi-algorithm

http://cpmarkchang.logdown.com/posts/192352

CRFs [Lafferty, Pereira, and McCallum 2001]

- Another sequence model: Conditional Random Fields (CRFs)
- A whole-sequence conditional model rather than a chaining of local models.

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

- The space of c' s is now the space of sequences
  - But if the features f_i remain local, the conditional sequence likelihood can be calculated exactly using dynamic programming
- Training is slower, but CRFs avoid causal-competition biases
- These (or a variant using a max margin criterion) are seen as the state-of-the-art these days … but in practice usually work much the same as MEMMs.

Maximum entropy sequence models

Maximum entropy Markov models (MEMMs) or Conditional Markov models