

# Chinese Language Understanding and Knowledge Graph

## 中文語言理解與知識圖譜

馬偉雲

中研院詞庫小組主持人

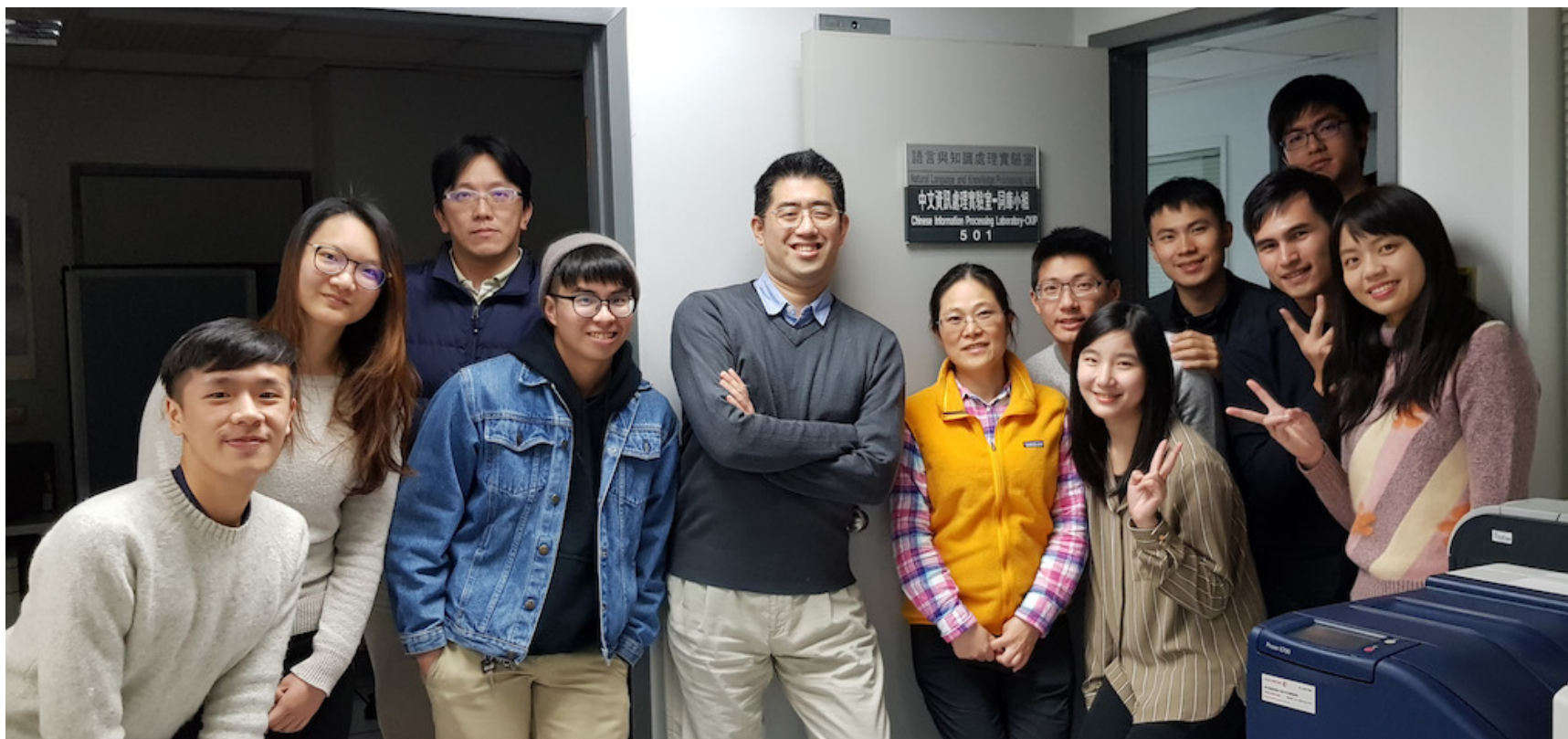
中研院資訊所助研究員

[ma@iis.sinica.edu.tw](mailto:ma@iis.sinica.edu.tw)

2018/05/25

# 中研院詞庫小組(CKIP)

- 中研院資訊所、語言所於民國七十五年成立一個跨所合作的中文計算語言研究小組，共同合作建構中文自然語言處理的資源與研究環境，為國內外中文自然語言處理及其相關研究提供基本的研究資料與知識架構。目前有五個主要研究方向：[深度學習](#)、[自然語言理解](#)、[知識表達](#)、[知識擷取](#)、[聊天機器人](#)。



# Outline

- **Part 1 Introduction to NLP**
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- **Part 2 NLP techniques**
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- **Part 3 NLP applications:**
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- **Part 4 Knowledge Graph – E-HowNet**

# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# What is Natural Language?

程式語言

自然語言



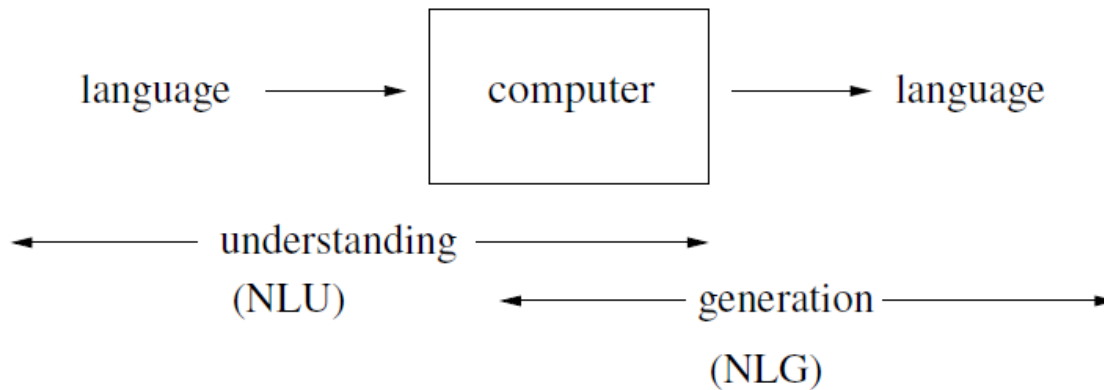
為計算機而設的語言  
依應用目的會使用不同程式語言  
例如 JavaScript, Python, PHP...



人類為了溝通而創造的語言  
有結構性語法 (Grammar)  
形式包含文字、語音、音樂等等

# What is Natural Language Processing (NLP)?

computers using natural language as input and/or output



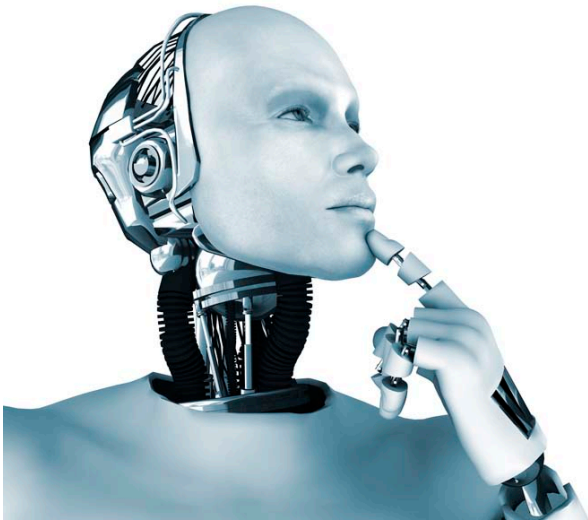
# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet



# How to teach a computer to learn language?

- 利用機器學習的演算法，讓電腦學會從訓練的原始資料或標記資料中，自動歸納出語言的特性。





# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# Chinese NLP by CKIP

- 中文斷詞系統 ([ckipsvr.iis.sinica.edu.tw](http://ckipsvr.iis.sinica.edu.tw))
- 中文剖析系統 ([parser.iis.sinica.edu.tw](http://parser.iis.sinica.edu.tw))
- 中文詞彙特性速描系統  
([wordsketch.ling.sinica.edu.tw](http://wordsketch.ling.sinica.edu.tw))
- 廣義知網線上系統 ([ehownet.iis.sinica.edu.tw](http://ehownet.iis.sinica.edu.tw))
- 輿情分析系統 ([learn.iis.sinica.edu.tw:9187](http://learn.iis.sinica.edu.tw:9187))
- 實體辨識系統 ([deep.iis.sinica.edu.tw:9001](http://deep.iis.sinica.edu.tw:9001))
- 聊天機器人  
([learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/](http://learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/))
- 中文詞彙庫 ([ckip.iis.sinica.edu.tw:8080/license/](http://ckip.iis.sinica.edu.tw:8080/license/))
- ...

# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# Chinese Word Segmentation

- 詞 (word)
  - 最小，有意義，且可以自由使用的語言單位。
- 斷詞 (word segmentation)
  - 將一個句子，根據其語意分解成合理的詞
  - 任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理
  - 中文不像英文有明顯的分隔符號 (空白符號) 可用以區分各別的詞
  - 一般的斷詞技術多利用詞典中收錄的詞來跟文本做比對，找出可能包含的詞

# Chinese Word Segmentation

- 存在的問題
  - 歧義的結果
    - 例如：
      - 當原子結合成分子時。
      - 可能的斷詞結果：
        - 當 原子 結合 成分 子時 。
        - 當 原子 結合成 分子 時 。
  - 未知詞(新詞)：不在詞典內的詞
    - 例如：
      - 莊駿凱、強投

# 詞類標記

- 詞類 (PoS, part-of-speech tag)
  - 因詞彙語意的不同造成一個詞有多重詞類
    - 這份研究(Na) 報告指出，
    - 赴美威斯康辛大學研究(VE) 一年。
  - 詳細說明可參考技術報告 ([中文詞類分析 93-05](#))
    - 細詞類有178個
    - 簡化詞類有46個 [[參考文件](#)]

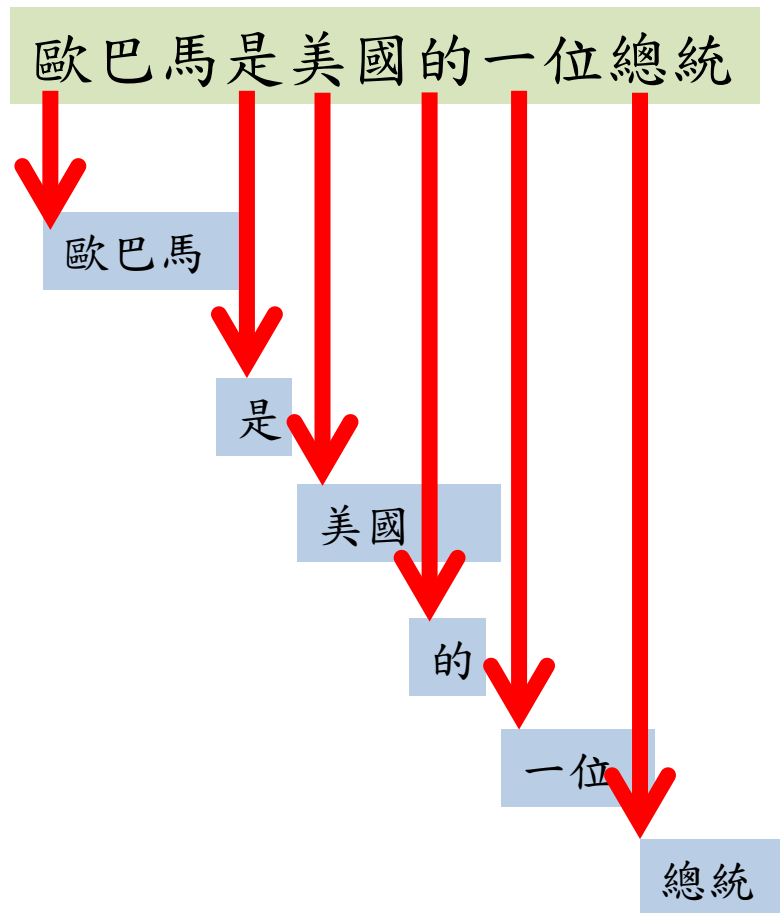


# Technique of Word Segmentation

- Word-level approach
- Character-level approach
  - Stanford
  - JiaBa
- Hybrid-level approach
  - CKIP Word Segmentation

# Word-level approach

- One technique is Maximum length Word First



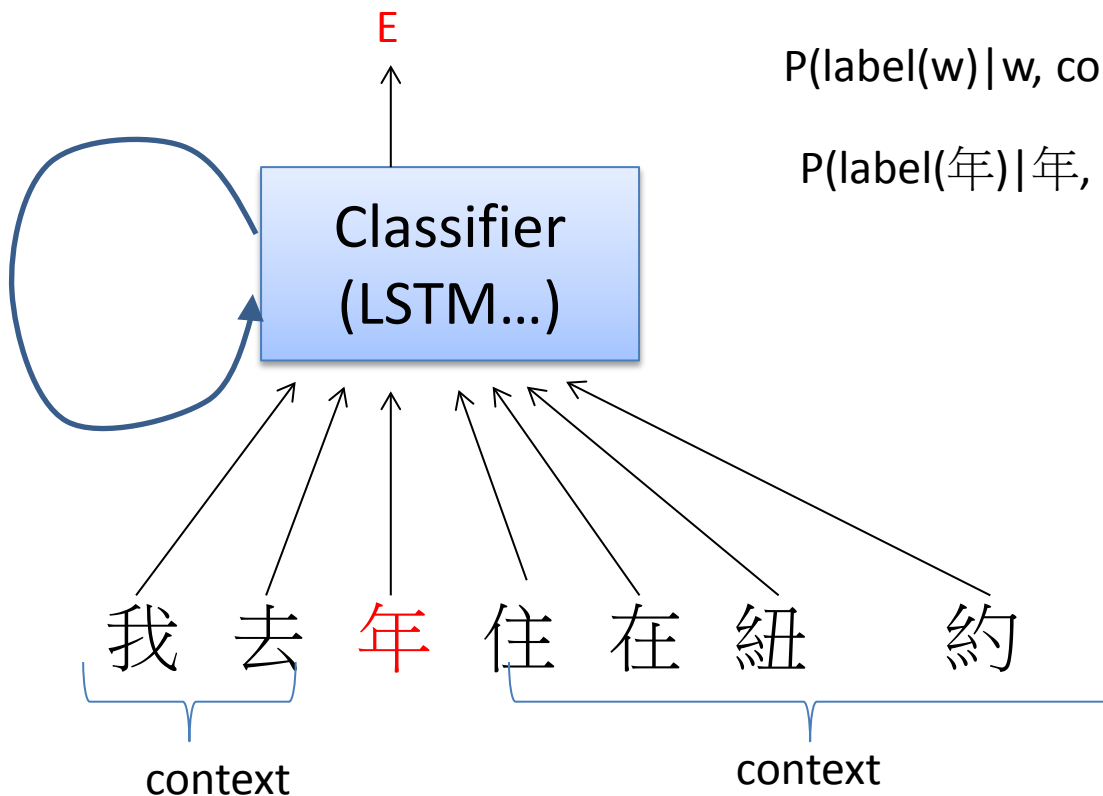
# Technique of Word Segmentation

- Word-level approach
- Character-level approach
  - Stanford
  - JiaBa
- Hybrid-level approach
  - CKIP Word Segmentation

# Character-level approach

- Character Sequence Labelling

我 去 年 住 在 紐 約      我 去 年 住 在 紐 約  
O   B   E   O   O   B   E



$P(\text{label}(w) | w, \text{context}(w), \text{previous words' labels})$

$P(\text{label}(\text{年}) | \text{年}, \text{context}(\text{年}), \text{label}(\text{去}), \text{label}(\text{我}))$

# Technique of Word Segmentation

- Word-level approach
- Character-level approach
  - Stanford
  - JiaBa
- Hybrid-level approach
  - CKIP Word Segmentation

# Technique of Word Segmentation

- Word-level approach
- Character-level approach
  - Stanford
  - JiaBa
- Hybrid-level approach
  - CKIP Word Segmentation

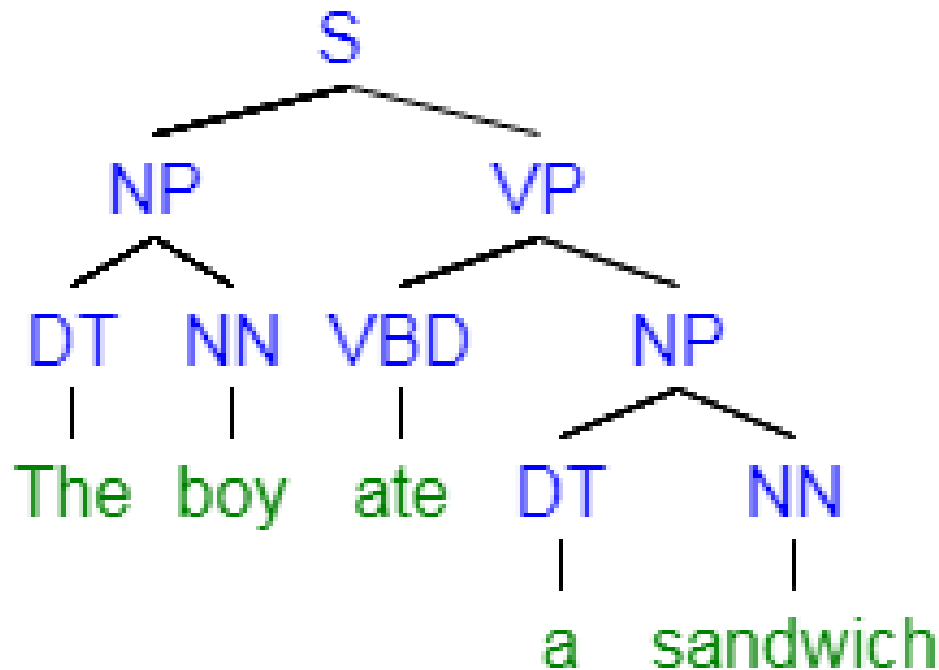


# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - [Syntactic Parsing](#)
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# Syntactic Structures by Syntactic Parsing

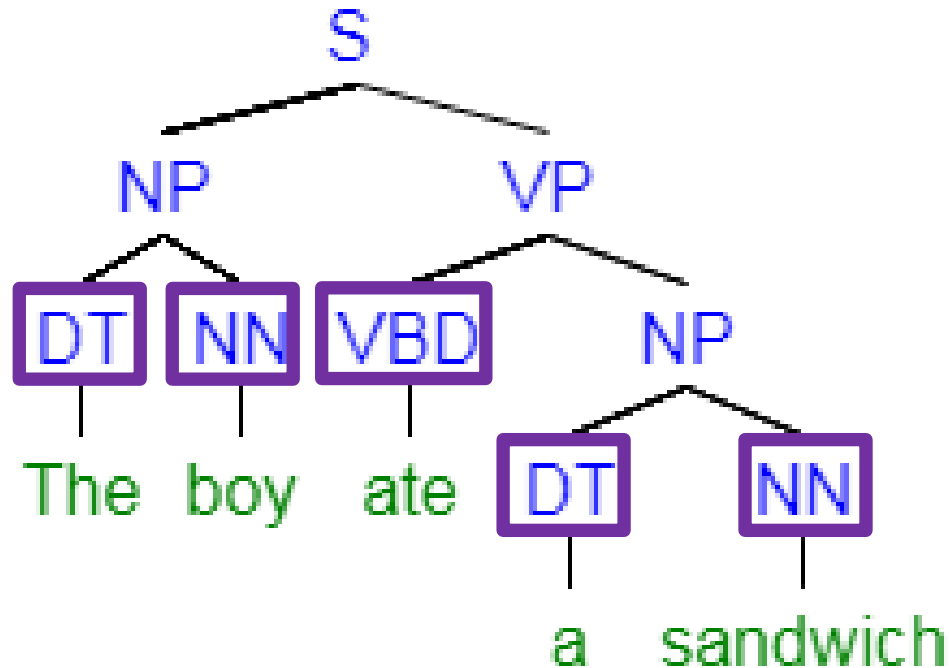
- Provide syntactic structure for a given sentence



# The Information Conveyed by Parse Trees

(1) Part of speech for each word

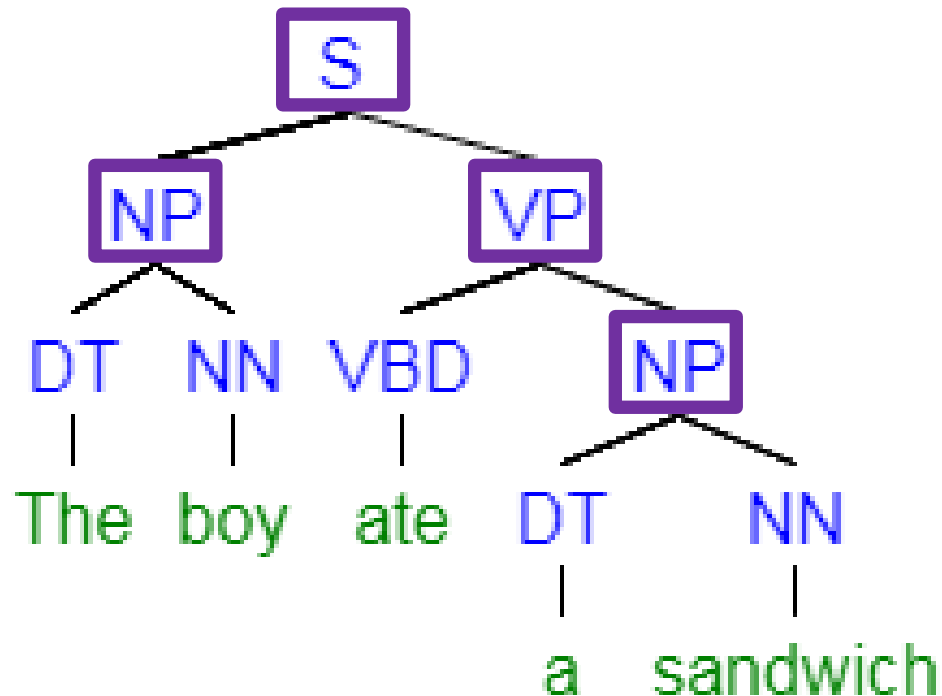
NN = noun, VBD = verb(past tense), DT = determiner)



# The Information Conveyed by Parse Trees

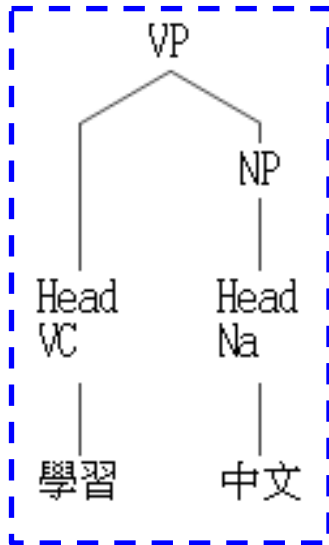
## (2) Phrases

NP = noun phrase, VP = verb phrase, S = sentence

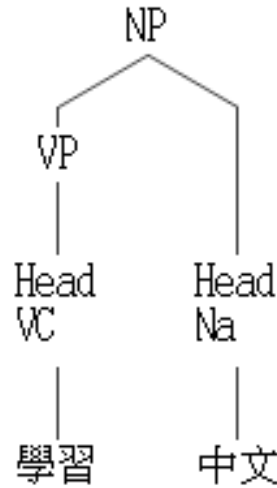


# Syntactic Ambiguity

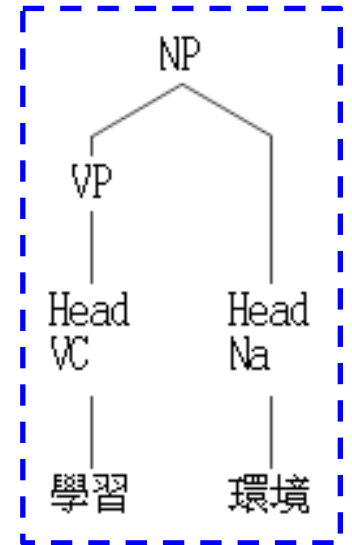
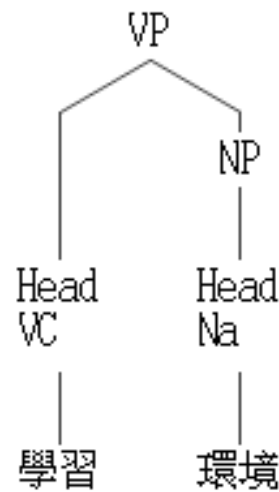
學習(Vt) 中文(N)



(0)

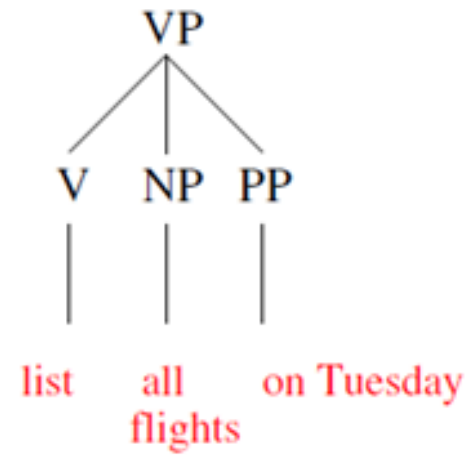
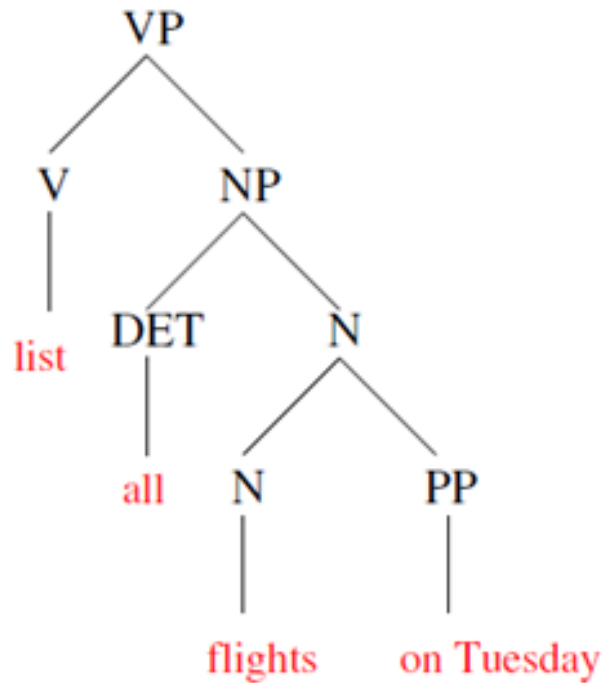


學習(Vt) 環境(N)



(0)

# Syntactic Ambiguity



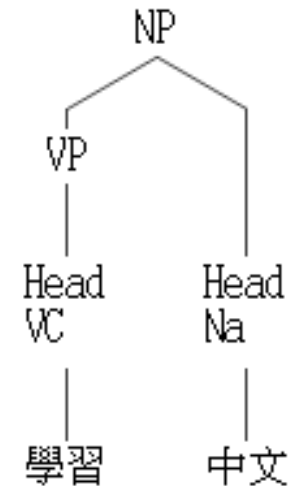
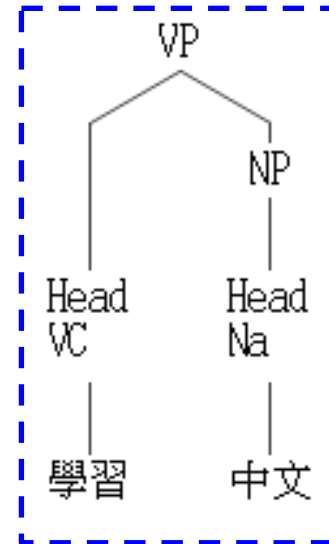
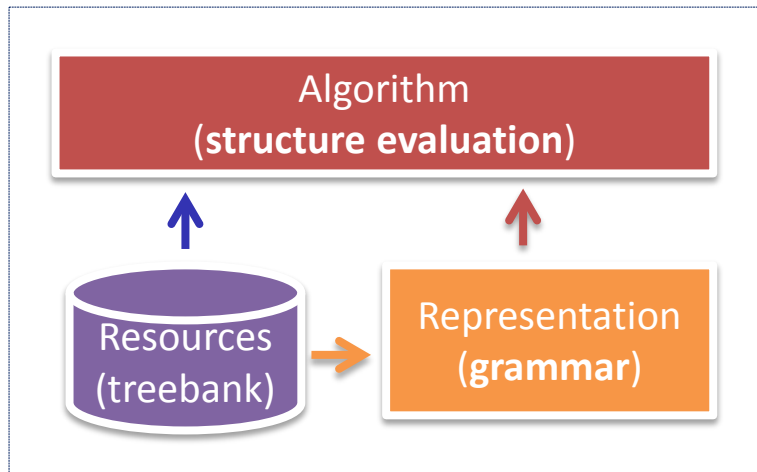


# Parsing = Structure Disambiguation

學習(Vt) 中文(N)



parser



(0)

# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - **Named Entity Recognition**
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# To Locate and Classify Named Entities (NEs)

*The defense secretary Donald Rumsfeld*

*ORG*                      *PER*

<u>NE Labels</u>
<i>PER (person)</i>
<i>ORG (organization)</i>

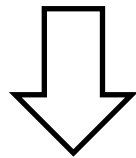
# Popular Approaches: Sequential Labeling NER (Seq-NER)

*The defense secretary Donald Rumsfeld*

*ORG PER*

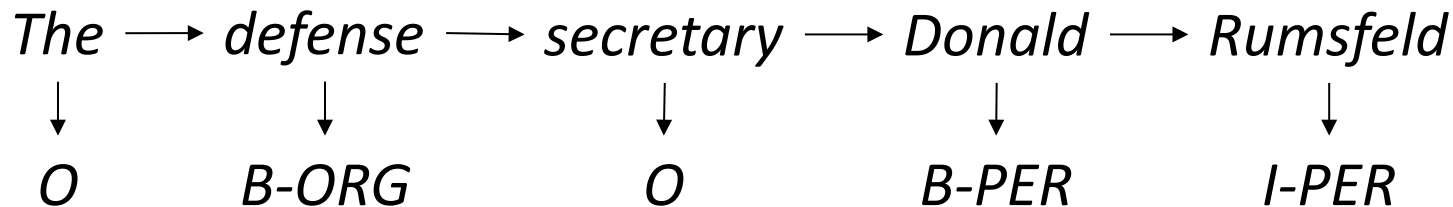


Seq Labels
<i>B (begin)</i>
<i>I (inside)</i>
<i>O (outside)</i>

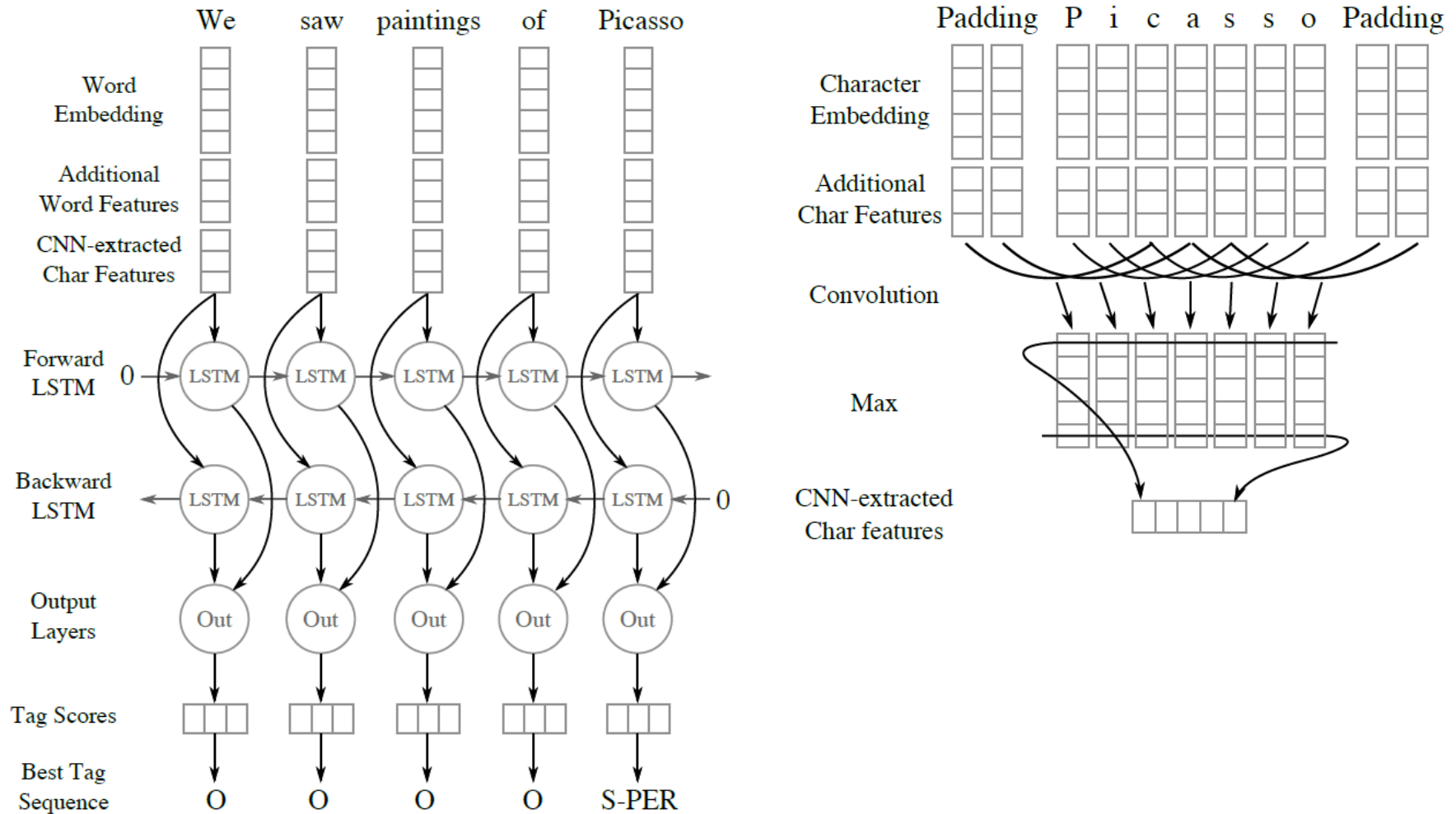


*The* → *defense* → *secretary* → *Donald* → *Rumsfeld*

*O*    *B-ORG*    *O*    *B-PER*    *I-PER*



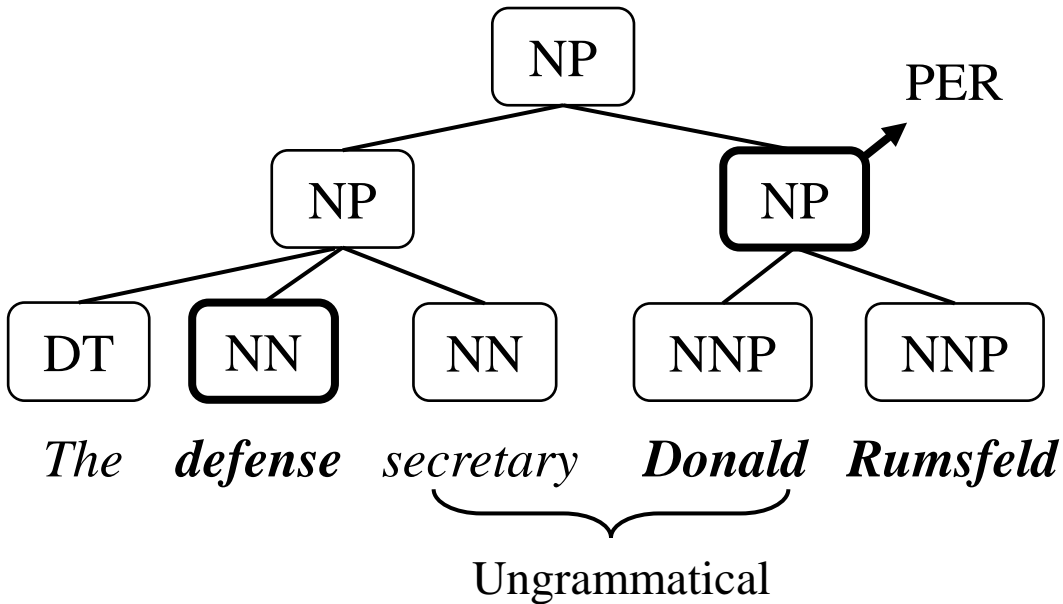
# Chiu and Nichols (2016)



J. P. Chiu and E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

# Motivation

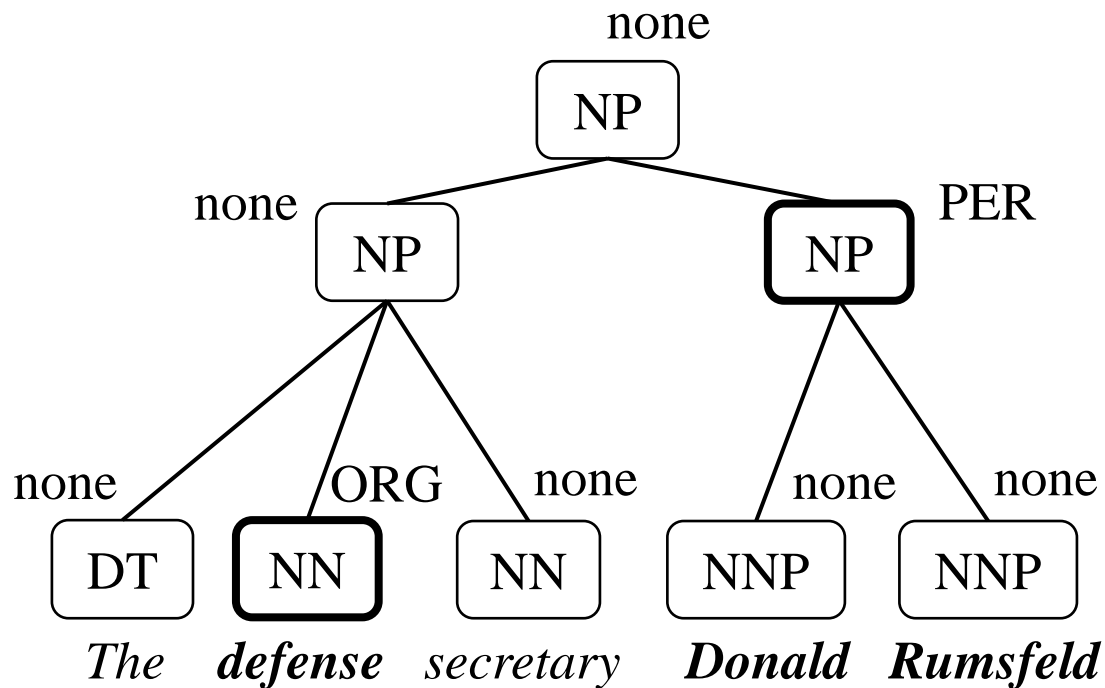
- Our observation
  - most NEs are constituents



Constituent → A plausible NE candidate

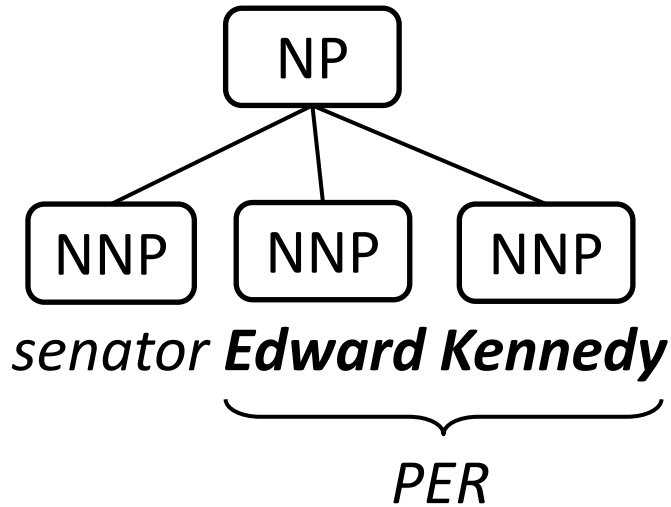
Ungrammatical → Unlikely an NE

# Our Idea: NER through labeling each constituent

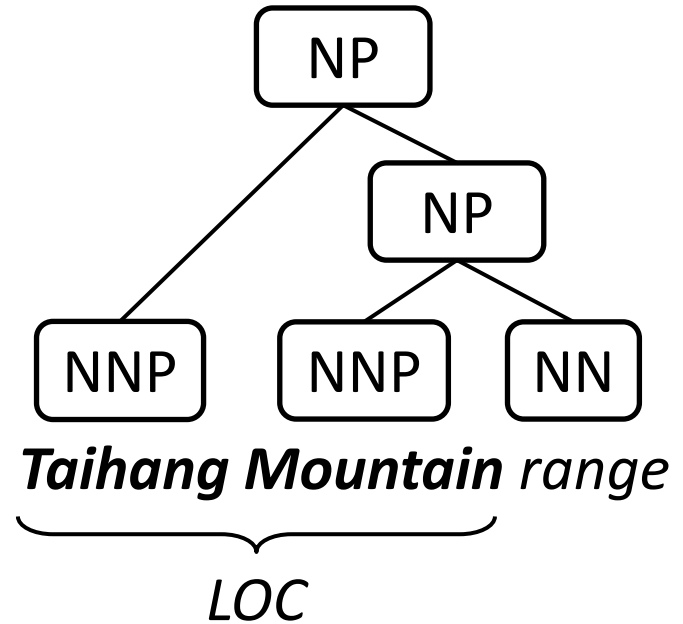


# But...

- There are still some inconsistent NE



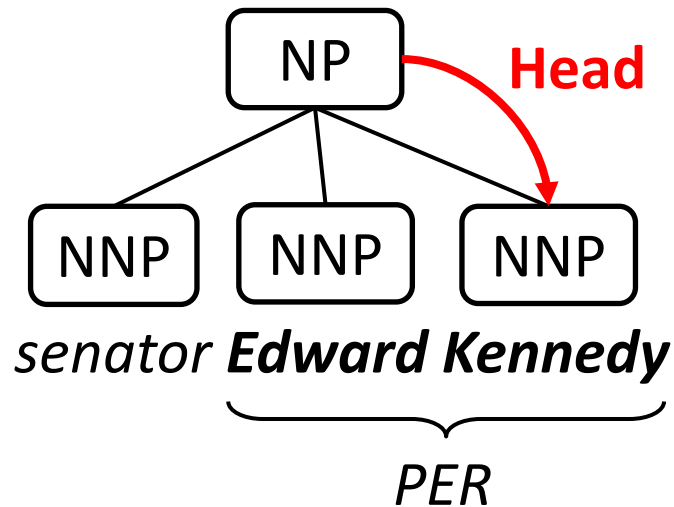
**Type-1**  
Cross Siblings



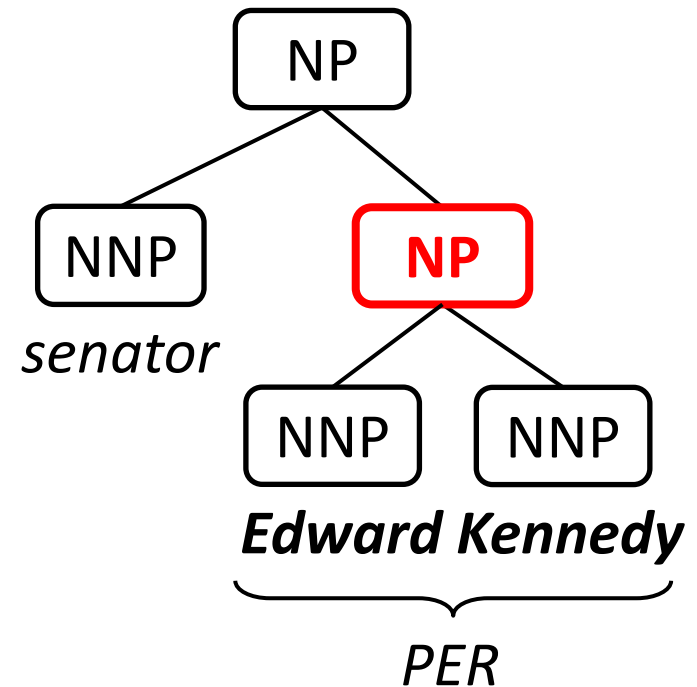
**Type-2**  
Cross Branches



# Eliminate Type-1: Constituency Tree Binarization

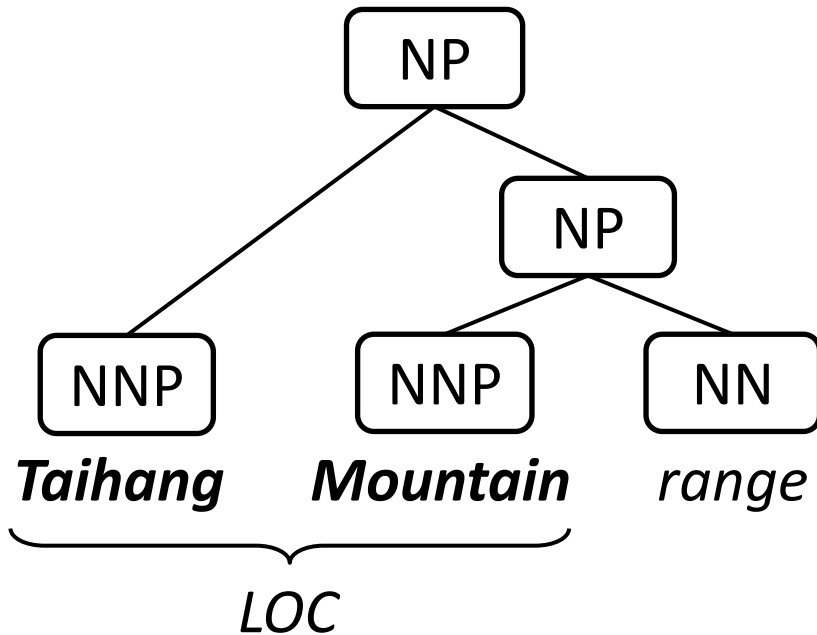


**Type-1**  
**Cross Siblings**

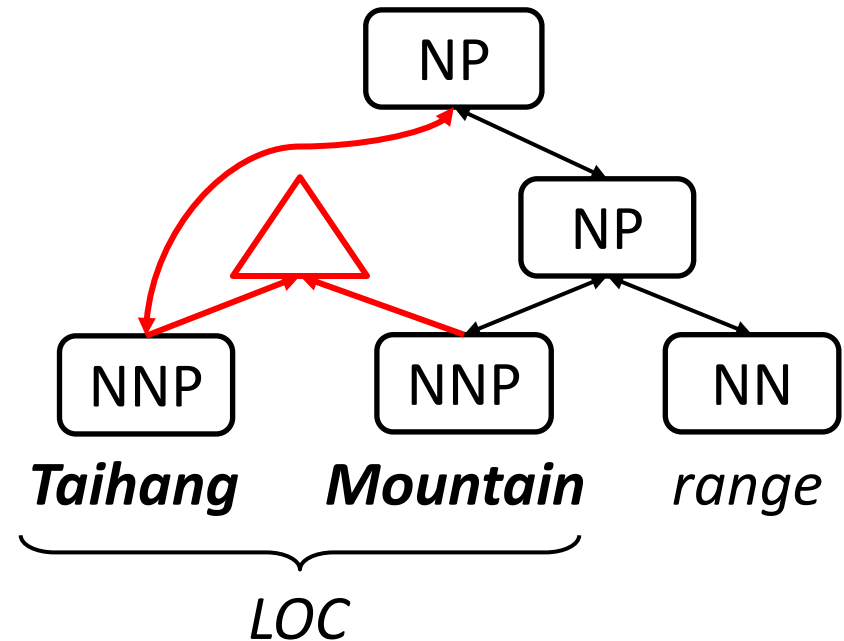


**Consistent**

# Eliminate Type-2: Pyramid Construction

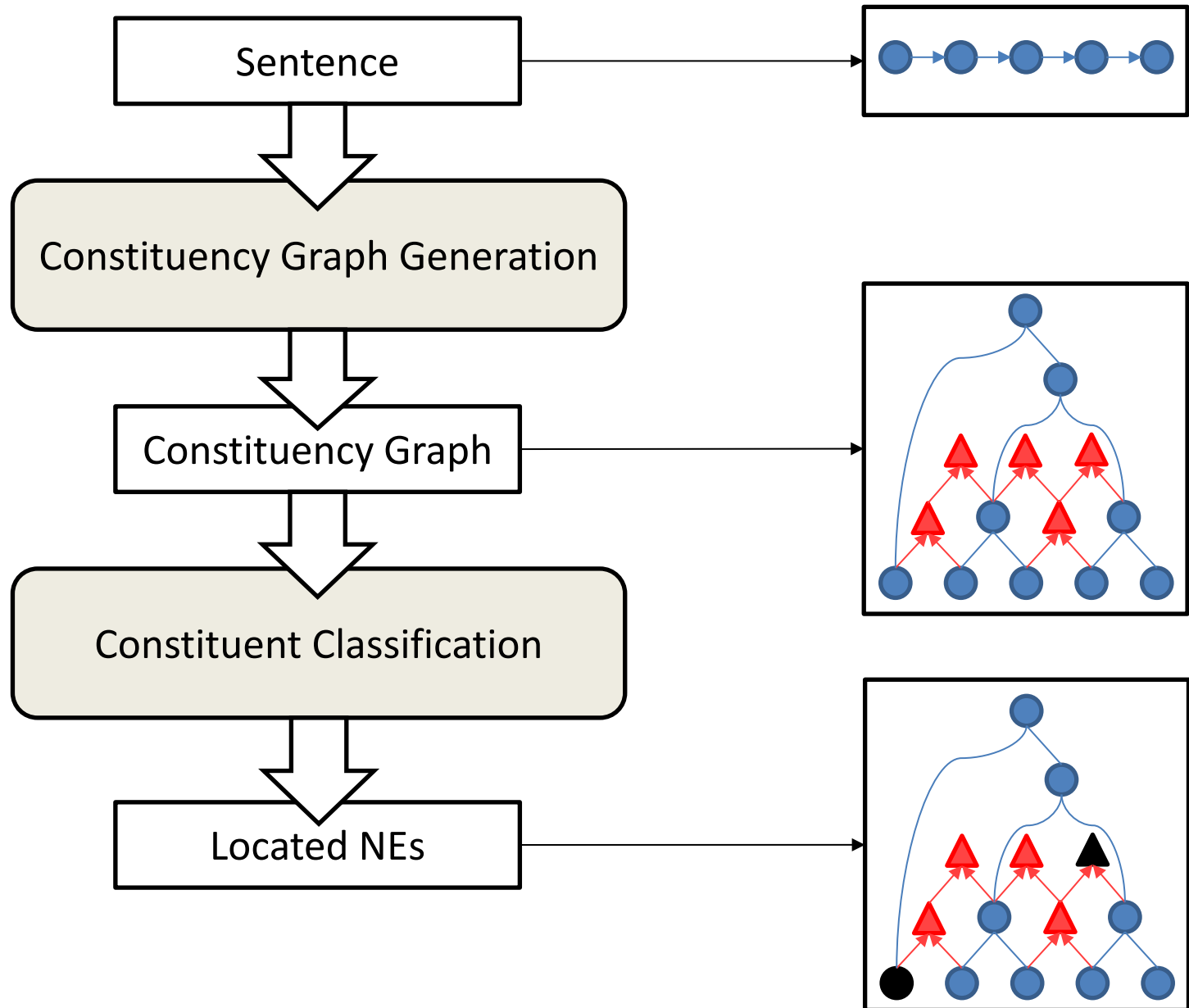


**Type-2**  
**Cross Branches**



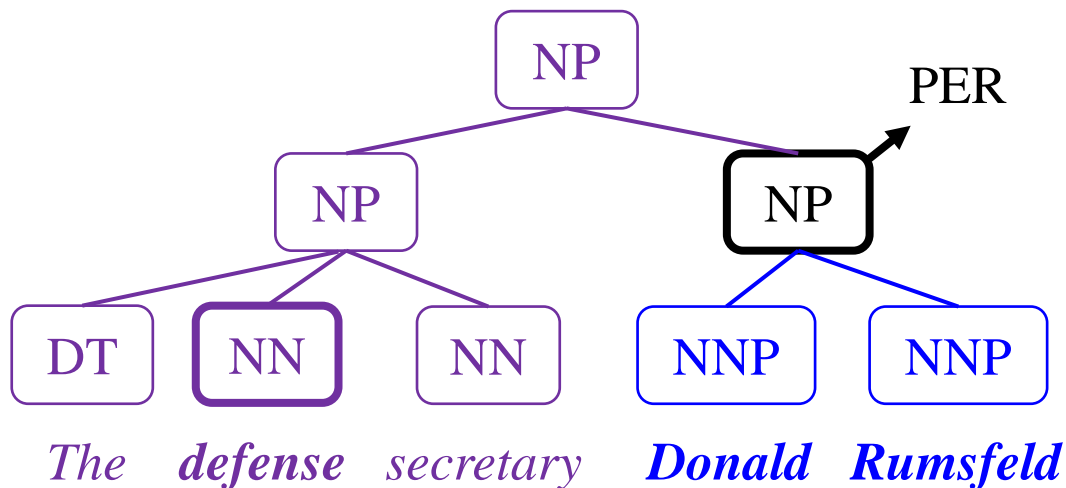
**No Inconsistencies**

# Constituency-Oriented NER

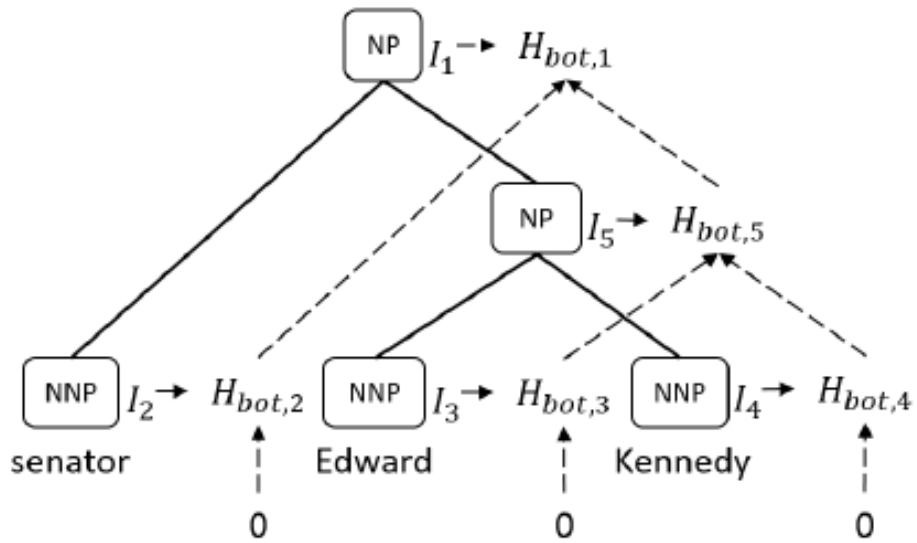


# Constituent Classification

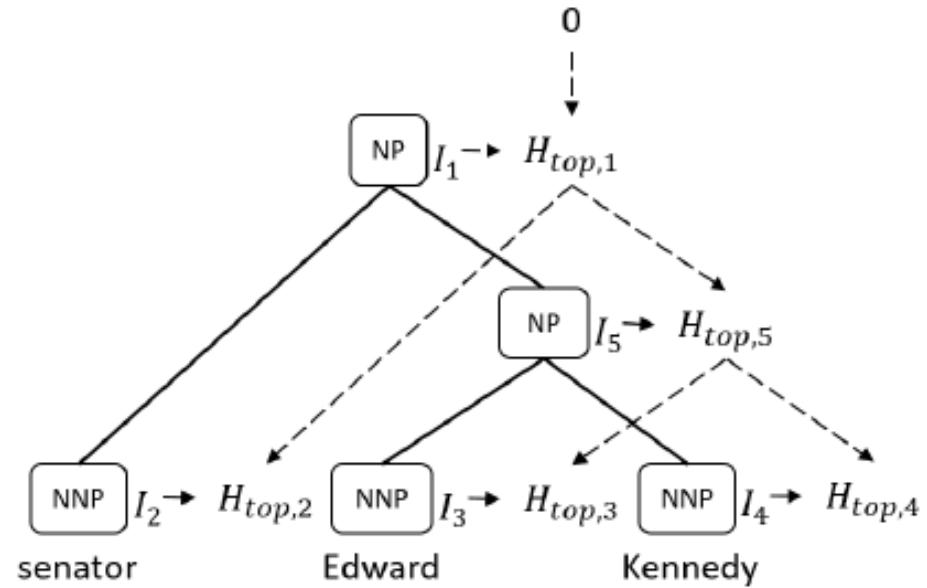
- Constituent Labelling should be using “**inside**” and “**outside**” information of the target constituent



### Bottom-Up



### Top-Down



# BRNN Output Layer

- For each node  $i$ , given
  - Left sibling  $l$
  - Right sibling  $r$
  - $H_x = H_{bot,x} + H_{top,x}$
- Compute
  - Predicted class probability distribution
$$O_i = \text{Softmax}((H_i || H_l || H_r) W_{out} + b_{out})$$

# Seq-Recurrent vs. Constituency-Oriented BRNN

**93% Consistency**

**97% Consistency**

<u>Model</u>	<u>CoNLL 2003</u>			<u>OntoNotes 5.0</u>		
	<u>Precision</u>	<u>Recall</u>	<u>F1</u>	<u>Precision</u>	<u>Recall</u>	<u>F1</u>
Bi-Recurrent	-	-	-	85.7	86.5	86.10
Chiu and Nichols (2016)	91.4	91.9	<b>91.62</b>	-	-	86.41
BRNN(-CNN)	90.2	87.7	88.91	88.0	86.5	<b>87.21</b>

Peng-Hsuan Li, Ruo-Ping Dong, Yu-Siang Wang, Ju-Chieh Chou and Wei-Yun Ma. 2017. Leveraging Linguistic Structures for Named Entity Recognition with Bidirectional Recursive Neural Networks. EMNLP 2017.

# Demo

- 中文斷詞系統 ([ckipsvr.iis.sinica.edu.tw](http://ckipsvr.iis.sinica.edu.tw))
- 中文剖析系統 ([parser.iis.sinica.edu.tw](http://parser.iis.sinica.edu.tw))
- 中文詞彙特性速描系統  
([wordsketch.ling.sinica.edu.tw](http://wordsketch.ling.sinica.edu.tw))
- 廣義知網線上系統 ([ehownet.iis.sinica.edu.tw](http://ehownet.iis.sinica.edu.tw))
- 輿情分析系統 ([learn.iis.sinica.edu.tw:9187](http://learn.iis.sinica.edu.tw:9187))
- [實體辨識系統 \(deep.iis.sinica.edu.tw:9001\)](http://deep.iis.sinica.edu.tw:9001)
- 聊天機器人  
([learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/](http://learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/))
- 中文詞彙庫 ([ckip.iis.sinica.edu.tw:8080/license/](http://ckip.iis.sinica.edu.tw:8080/license/))
- ...



# Outline

- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet

# 背景

- 大數據時代來臨，如何自動從眾多資料中抽取出有意義的訊息、分析市場、探勘輿情，對許多政府單位、企業而言是相當重要的應用。
  - 選戰-2014年的台北市長選舉一役 (協助柯P蒐集選情輿論)、歐巴馬、川普
  - 精準行銷
    - 統計汽車經常會和什麼主題放在一起討論，如此一來，車商就可以知道應該把廣告下在什麼樣的地方。
  - 公關危機化解
    - 即時掌握網路輿情，化解可能的公關危機。

# 時下一般的輿情分析系統

- 大多**Online**系統會提供三種面向的分析
  - 聲量分析：不同的品牌與商品在網路上討論的熱度
  - 趨勢分析：目前網路上最熱門、最新的討論內容與趨勢。
  - 情感分析：判斷消費者在討論品牌與商品時的情緒與喜好

## 百大口碑 / 人物

分析期間：2017/04/30~2017/10/26

## 立法委員

1  黃國昌

正面23% 中立50% 負面27%

71,093 篇 [MORE >](#)

2  柯建銘

正面22% 中立39% 負面39%

23,019 篇 [MORE >](#)

3  徐永明

正面19% 中立42% 負面39%


21,811 篇 [MORE >](#)

## 政治人物

1  柯文哲

正面20% 中立48% 負面32%

738,441 篇 [MORE >](#)

2  蔡英文

正面26% 中立43% 負面31%

541,799 篇 [MORE >](#)

3  賴清德

正面24% 中立53% 負面23%

318,101 篇 [MORE >](#)

## 通告藝人

1  白雲

正面19% 中立51% 負面30%

41,164 篇 [MORE >](#)

2  從從 唐從聖

正面15% 中立50% 負面35%

39,306 篇 [MORE >](#)

3  解婕翎

正面28% 中立54% 負面18%

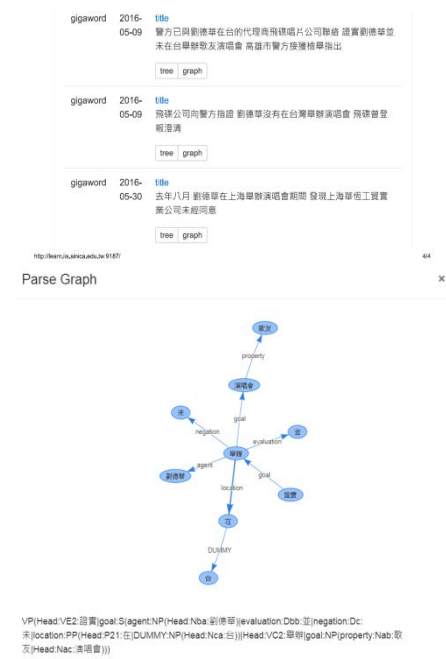
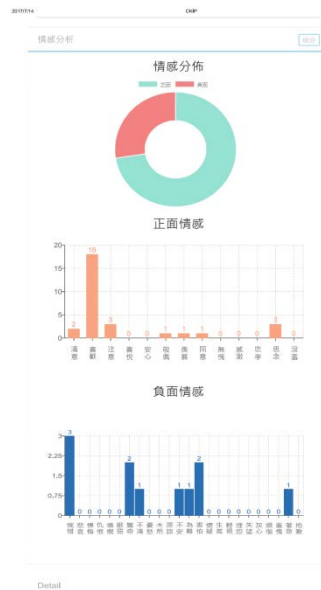
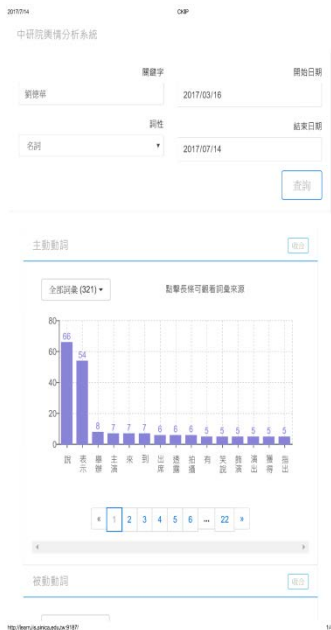
12,513 篇 [MORE >](#)

# 中研院所開發的輿情分析系統

- 聲量分析
- 趨勢分析
- 情感分析：除了正負向之外, 另有**50**類左右  
細緻情緒分類
- 細緻的事件與語義分析
- 可追蹤的分析來源, 並將分析結果視覺化

# 中研院輿情分析系統

- <http://learn.iis.sinica.edu.tw:9187/>
  - Deep Semantic Analysis
  - Comprehensive Sentiment Analysis
  - Ability to Trace Every Evidence



# Outline

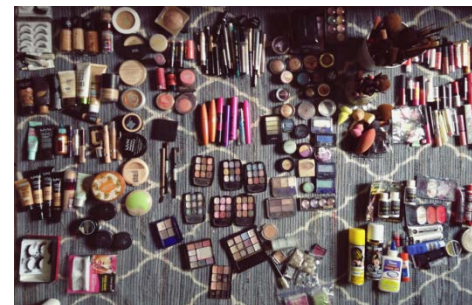
- Part 1 Introduction to NLP
  - What is Natural Language Processing (NLP)?
  - How to teach a computer to learn language?
  - Chinese NLP by CKIP
- Part 2 NLP techniques
  - Chinese Word Segmentation
  - Syntactic Parsing
  - Named Entity Recognition
- Part 3 NLP applications:
  - 中研院輿情分析系統
  - 美妝保養聊天機器人
- Part 4 Knowledge Graph – E-HowNet



PerMu  
By CKIP



- **PerMu**是一款中研院詞庫小組開發的美妝保養聊天機器人，是個人化的美妝保養顧問，**PerMu**榮獲**PIXNET 2017** 聊天機器人大賽的冠軍(最佳產品獎) 已在**Facebook**上對外開放。
- 線上使用網址：  
<https://www.facebook.com/PerMu-133307777270004/>

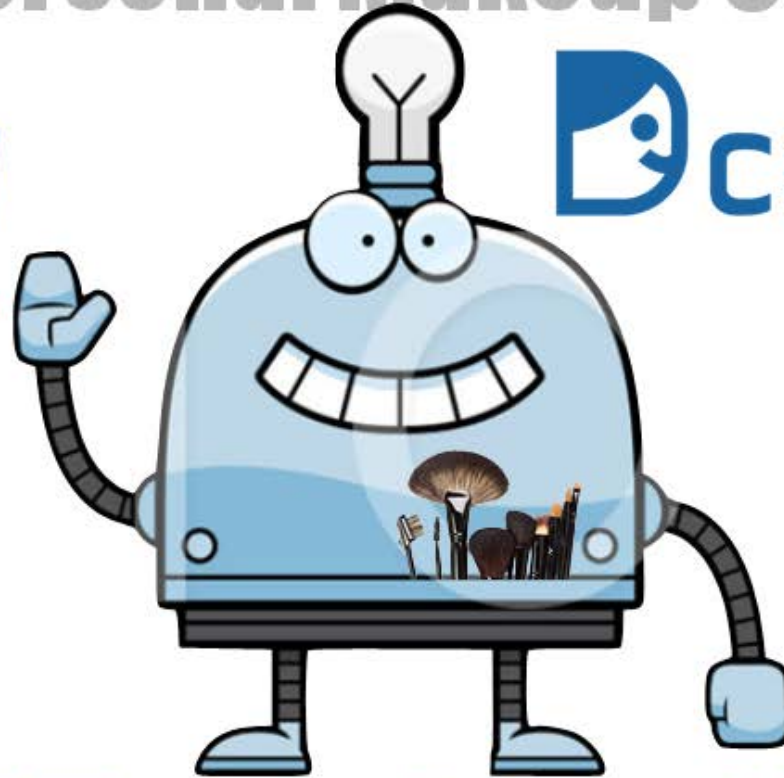


**Perμ** Personal Makeup Consultant

PIXNET

styleMe

card



EHOWNET

w2v

NER

IR



# Recommend Items

能推薦粉底給我嗎？



要不要試試亮膚色的產品？

好啊



要不要試試INTEGRATE 絕色魅癮的柔焦輕透美肌粉餅SPF16/PA++ 明亮膚色？價錢約在370元（10g）左右～

有貴一點的嗎？



要不要試試專櫃的產品？

好的



要不要試試LANCÔME 蘭蔻的水感奇蹟粉餅 PO-01？價錢約在1450元（10g）左右～這個商品獨家AURA 光感科技，能提高肌膚光線反射度，立即明亮膚色。添加玫瑰精萃給予肌膚服貼保濕妝感，鎖住肌膚水分，輕透妝感，立即柔焦毛孔，重複上妝也不厚粉！



要不要試試專櫃的產品？



要不要試試酒紅色的產品？

想找唇膏

好啊

我想找霧面偏橘色的唇膏



ettusais 艾杜紗 糖果啾啾護唇精華棒 OR 橘色 2.2g 420

雅詩





## More Real Product Effects

我想要找眼線筆

要不要試試heme的速描防水眼線眼影筆-咖啡？  
價錢約在149元 (1.5g) 左右~這個商品質地為  
柔順慕絲狀，讓你使用順暢不傷眼皮。帶有最適  
合亞洲膚色的細緻珠光，不論單獨上色或是漸層  
混搭皆能打造奪目眼妝，獨家德國科技的防水薄  
膜可長效防水，使用一整天不暈染。高密度棉棒  
的雙頭筆管的設計，輕鬆讓眼線化身眼影。



防水效果如何？

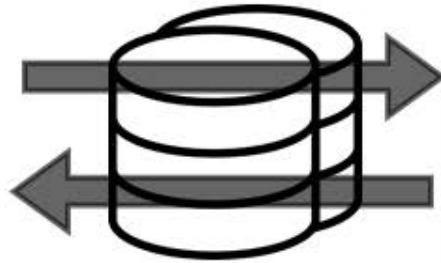
畫完馬上推開就能變成眼影  
我測試過防水效果  
稍等它乾掉之後沖水都不會掉  
(<http://mita1229.pixnet.net/blog/post/220217825>)



# NER+IE



# Name Entity Linking







## More Product Info

想找唇膏

要不要試試SOFINA 蘇菲娜的星鑽美形輕潤持色唇彩？價錢約在900元（3.8g）左右～這個商品水潤色澤持續配方及Ceramide精華等高保濕成分，經過長時間、用餐，雙唇依舊像剛擦上唇彩般滋潤顯色，長時間持續飽和色澤，搭配上獨特研發的「金色柔彩珍珠」，自然細緻不泛白光。



有更多介紹嗎？

水潤色澤持續配方及Ceramide精華等高保濕成分，用餐時間1hr，雖然剛擦上雙唇，色澤還能維持很久，這款獨特添加了粉紅珠光澤粉體，呈現的效果自然細緻。



w2v +  
Sent Emb





## Product Figures

要不要試試KATE 凱婷的流光潤彩唇膏 RD-2？價錢約在300元（2.5g）左右～這個商品綻放瑰麗妍色，飾出亮澤水潤理想俏唇的高顯色唇膏，柔滑觸感，完美貼合雙唇，全10色。



有圖嗎



這是KATE 凱婷的流光潤彩唇膏 RD-2的產品圖！

我想要找遮瑕

要不要試試CANMAKE的全方位遮瑕組SPF50 PA++++？價錢約在385元（3.9g）左右～這個商品三種顏色，只要一盒，就能完美遮蓋各種肌膚瑕疵，呈現健康美肌，打造出清爽妝感。



有圖嗎？





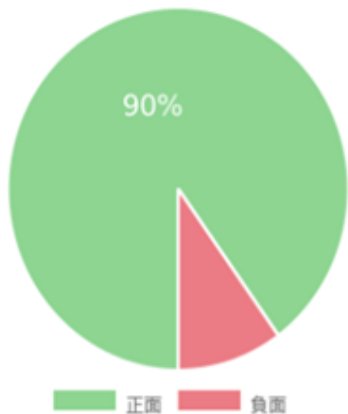
# Product Rating

要不要試試Dior 迪奧的藍星唇膏？價錢約在1100元 (3.5g) 左右～這個商品豐潤飽和的色澤與質地，歸功於天然保養成分「海茴香萃取物」，能修護並緊實雙唇同時提供絕佳的保濕效果；「玻尿酸微珠」可以立即撫平唇紋、豐盈雙唇，展現絲滑亮澤質感。

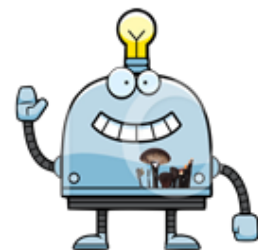


你覺得好用嗎

討論熱度：12



根據我的分析～使用者對於 Dior 迪奧的藍星唇膏 的反應







# Links to Products

能介紹粉底嗎~



要不要試試專櫃的產品？

好的



要不要試試白皙膚色的產品？

好



要不要試試LANCÔME 蘭蔻的水感奇蹟粉餅 O-01？價錢約在1450元（10g）左右~這個商品獨家AURA光感科技，能提高肌膚光線反射度，立即明亮膚色。添加玫瑰精萃給予肌膚服貼保濕妝感，鎖住肌膚水分，輕透妝感，立即柔焦毛孔，重複上妝也不厚粉！

哪裡買呢~

<http://www.lancome.com.tw/彩妝/臉部妝容/底妝/HK5863400-LAC.html#q=水感奇蹟&start=2>

這個商品真的很不錯，獨家AURA光感科技，能提高肌膚光線反射度，增加細胞感光體幫助肌膚光線均勻折射產生柔和光感，添加玫瑰精萃給予肌膚服貼保濕妝感，給予透亮水感的成分呢，輕透妝感，立即柔焦毛孔，重複上妝也不厚粉。



保養 · 彩妝 · 頂級保養 · 香水 · 線上特惠組 · 線上購物優惠 · 品牌故事 · 新品上市 · 銷售排行

首頁 > 彩妝 > 臉部妝容 > 底妝 > 水感奇蹟玫瑰水粉餅(不含粉盒)

## 水感奇蹟玫瑰水粉餅蕊

TEINT MIRACLE COMPACT (SPF20 / PA++)



注入頂級玫瑰精華與玻尿酸，保濕力提升40%，高單位玻尿酸形成強力保濕膜，將玫瑰精華鎖在肌膚，上妝就像敷面膜，肌膚散光又透亮；結合最新科技「氣態粉體」，輕潔如空氣的極細粉體，確保粉體的1/10重量，一抹立即融合肌膚，完美修飾毛孔又零一。

了解更多

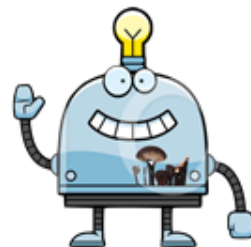


請來所有4色調

NT\$1,600 項目 只有3 還有現貨

加入購物車

加入願望清單





## Links to Tutorials and Info

我想學化妝



想要試試夜店的妝容嗎？

好



也許韓國姊姊寫的『韓國演員“高熙俊”時尚夜店妝！』適合你噢



[全文連結在這裡～](#)



也許韓國姊姊寫的『韓國演員“高熙俊”時尚夜店妝！』適合你噢

[全文連結在這裡～](#)

他的口紅是什麼顏色

橘紅色



我個人覺得她用的產品效果還不錯，也許可以入手一隻

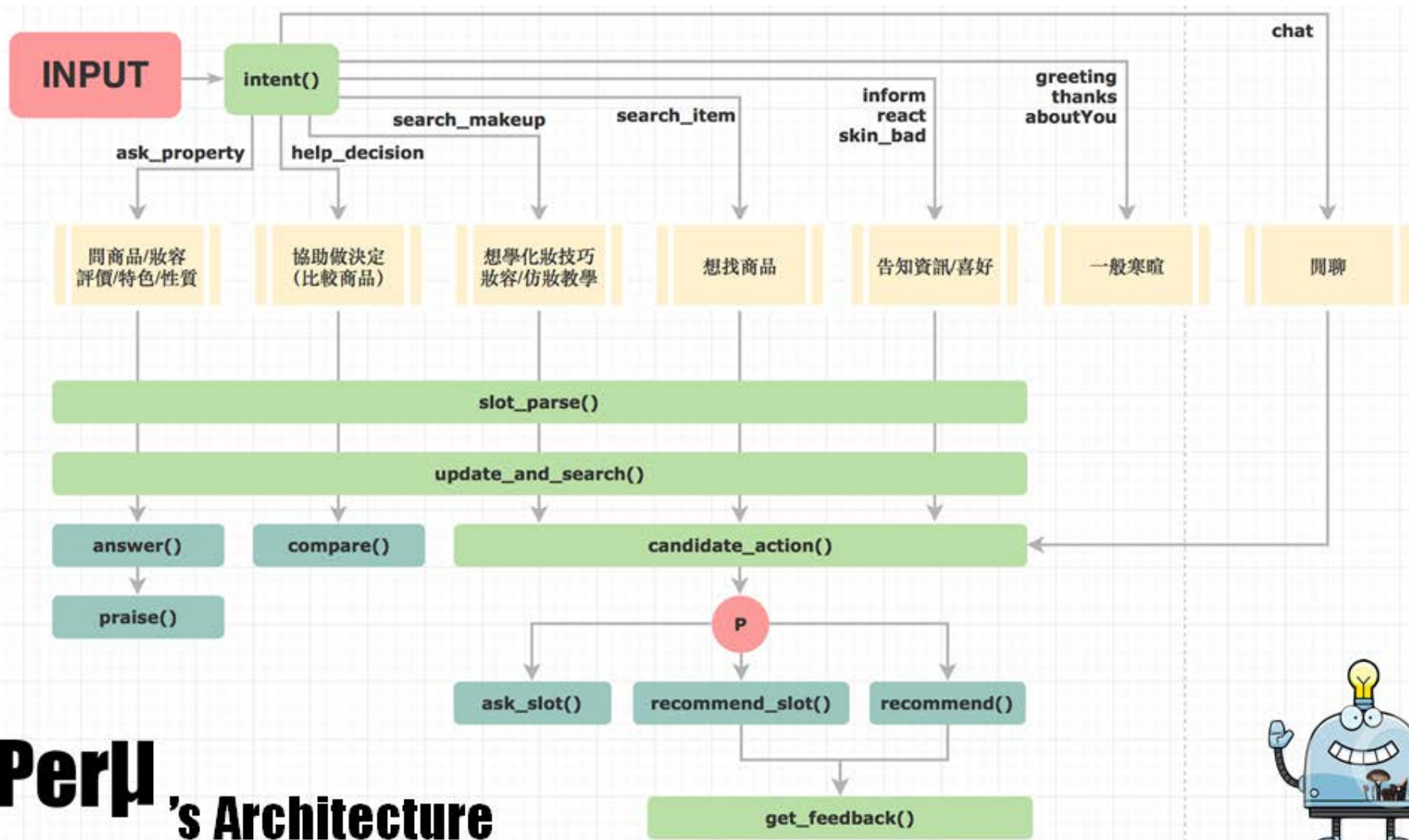
口紅是什麼牌子

M.A.C

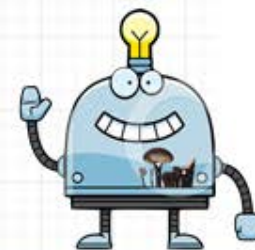


我個人覺得她用的產品效果還不錯，也許可以入手一隻





# Perμ's Architecture



# Outline

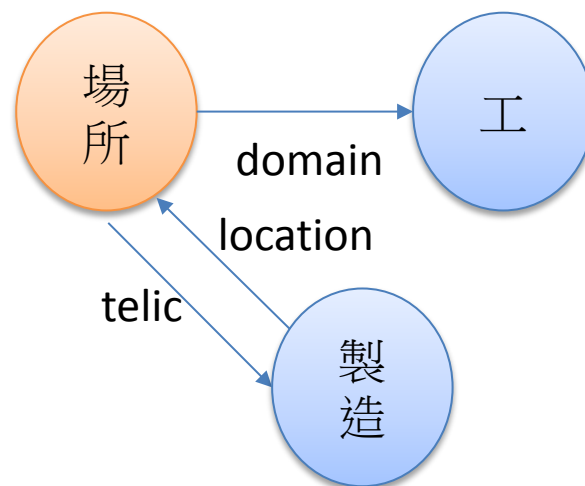
- Chinese NLU by CKIP
- Syntactic Structures by Syntactic Parsing
- Named Entity Recognition with Syntactic Structures
- Knowledge Graph for Chinese Common Sense
  - E-HowNet (廣義知網)

# 廣義知網(E-HowNet)

- 董振東先生於1988年左右創立知網。廣義知網承續知網(HowNet)的語意定義機制，將中央研究院詞庫小組辭典中的九萬多詞條與知網連結，以**通用的概念**為描述對象，建立並描述這些概念之間的**關係**。

Ex:工廠 def:{InstitutePlace|場所:domain={industrial|工},  
telic={produce|製造:location={~}}}

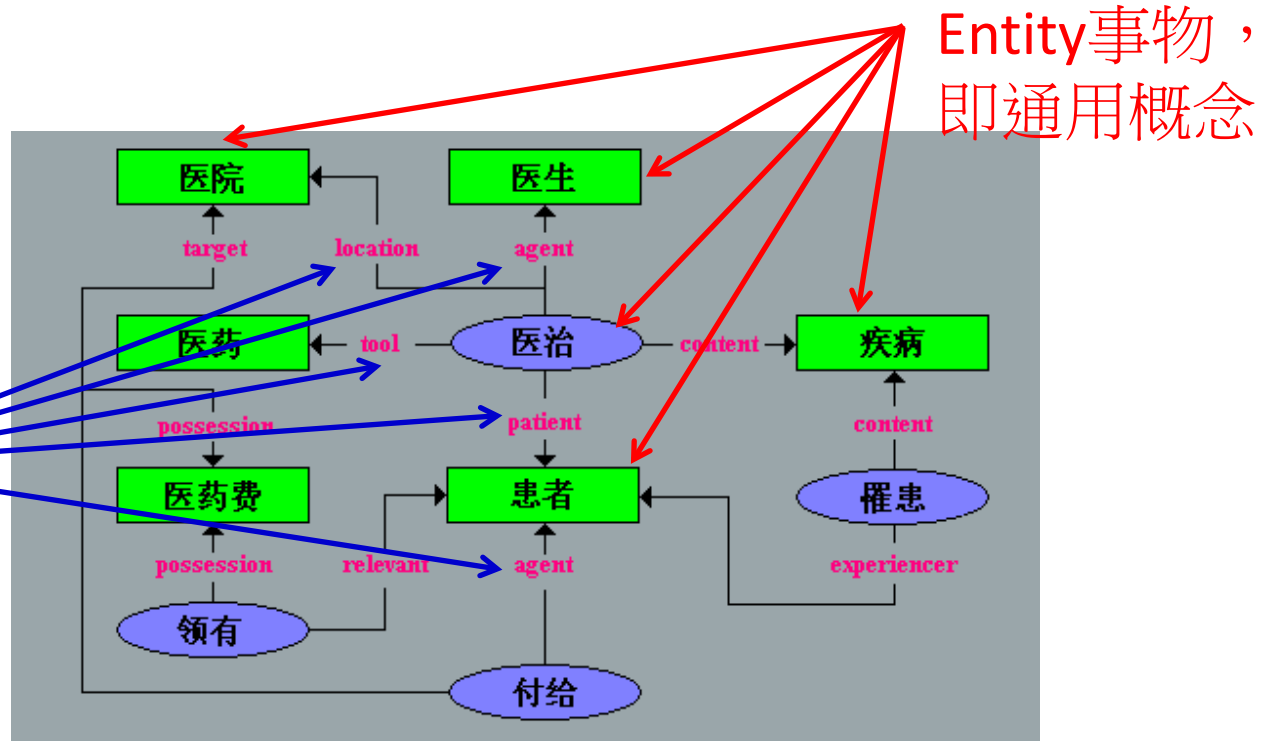
- 廣義知網(E-HowNet) 的特色
  - 繁體中文(九萬多詞條)
  - 功能詞
  - 多層次定義式
  - .....



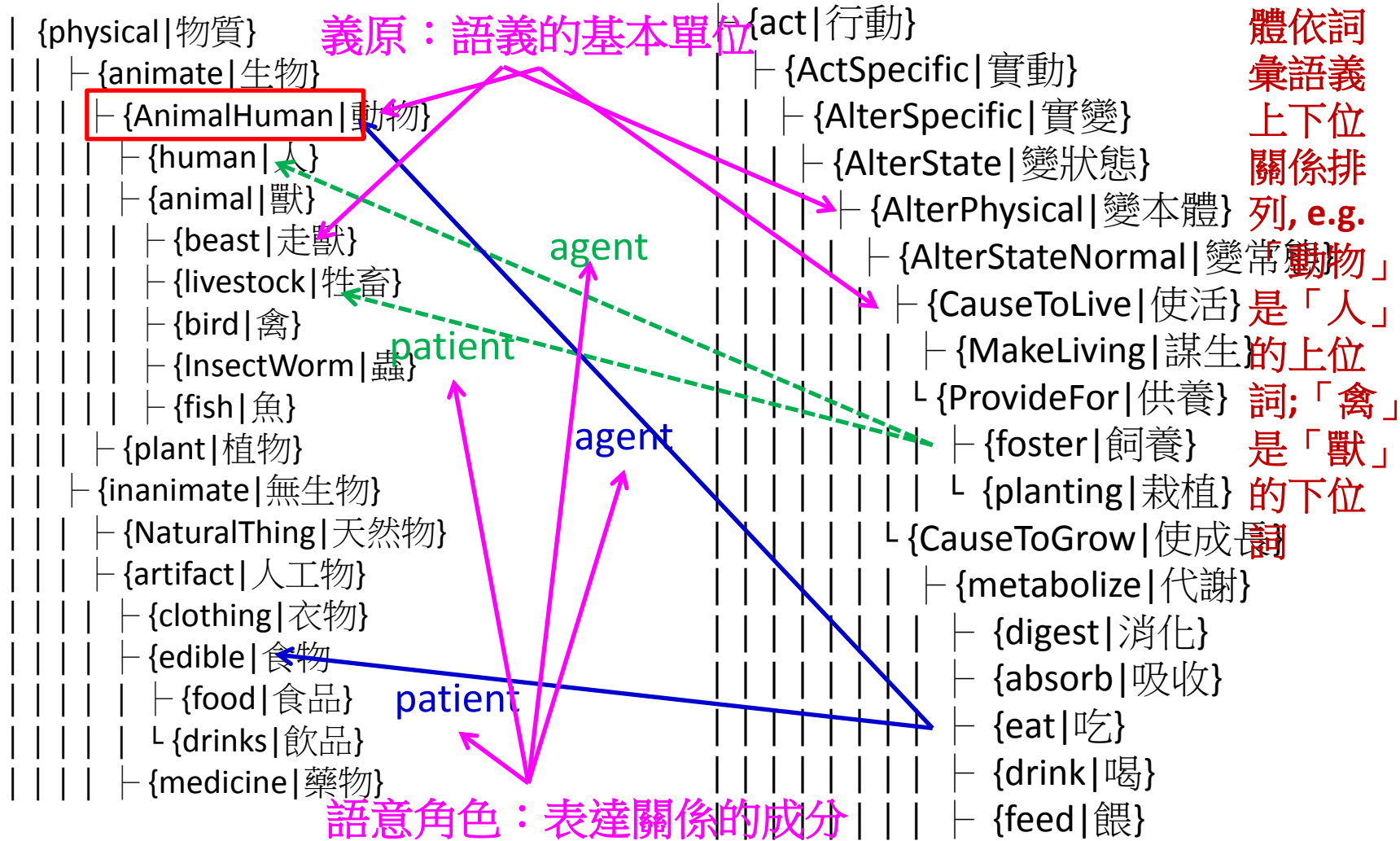
# 例子一患者

- 國語辭典：病人
- 廣義知網：{human|人:predication={or({ill|病態:theme={~}},{doctor|醫治:patient={~}})}}}

Relation 關  
聯



# E-HowNet知識本體(Ontology)



# 建構知識網絡

- 醫治 def:{doctor|醫治}
- 醫師 def:{human|人:  
domain={medical|醫},  
predication={doctor|醫治:  
agent={~}}}
- 醫院 def:{InstitutePlace|場所:  
domain={medical|醫},  
telic={doctor|醫治:  
content={disease|疾病},  
location={~}}}
- 醫藥罔效 def:{BeRecovered|復原:  
theme={disease|疾病},  
ability={least|無}}
- 醫藥費 def:{expenditure|費用:  
predication={doctor|醫治:  
price={~}}}

1. 用義原(entity)及語意角色(relation)來定義詞彙

2. 藉由詞彙的語義表達式，同時也建構了知識網絡，如醫治與疾病，疾病與復原.....等等的關係



# 構建同(近)義詞集

- 學生 def:{human|人:  
predication={study|學習:  
domain={education|教育},  
agent={~}}}  
**1. 相同的語義會有相同的表達式**
- 生員,生徒,弟子,門下,門徒,後學,徒子,徒孫,徒弟,桃李,莘莘學子,學子,學徒,學徒工,生...  
**2. 近似的語義會有近似的表達式**
- 留學生 def:{human|人:predication={study|學習:agent={~},location={foreign|外國},domain={education|教育}}}  
**3. 近義詞語義往往是修飾成分有所不同，具體表現在語意角色上**
- 病危 def:{ill|病態:manner={serious|嚴重}}  
久病 def:{ill|病態:duration={TimeLong|長時間}}  
累病 def:{ill|病態:cause={tired|疲乏}}  
抱病 def:{ill|病態:aspect={Vgoingon|進展}}  
復發 def:{ill|病態:frequency={again|再}}

# 多層次表達

## 西瓜

- 基本概念式
- 展開為義原表達式

def1: {西瓜 | watermelon}

def2: {fruit | 水果:

### 1. 義原：{英文|中文}

predication={contain | 包含:

content={liquid | 液:

quantity={many | 多}},

theme={~}}}

## 獅子狗

- def1: {狗 | dog: source={北京 | Beijing}}

- def2: {livestock | 牲畜:

telic={看家 | MindTheHouse:

agent={~}},

source={北京 | Beijing}}

- def3: {livestock | 牲畜: telic={TakeCare | 照料:

patient={family | 家庭},

agent={~}},

source = {capital | 國都:

name={"北京"},

location={China | 中國},

quantifier={definite | 定指}}}

### 2. 基本概念：{中文|英文}

基本概念保留更多的訊息

讓表達式簡化、清晰

### 3. 多層次的語義表達式

# Chinese NLU by CKIP

- 中文斷詞系統 ([ckipsvr.iis.sinica.edu.tw](http://ckipsvr.iis.sinica.edu.tw))
- 中文剖析系統 ([parser.iis.sinica.edu.tw](http://parser.iis.sinica.edu.tw))
- 中文詞彙特性速描系統  
([wordsketch.ling.sinica.edu.tw](http://wordsketch.ling.sinica.edu.tw))
- [廣義知網線上系統 \(ehownet.iis.sinica.edu.tw\)](http://ehownet.iis.sinica.edu.tw)
- 輿情分析系統 ([learn.iis.sinica.edu.tw:9187](http://learn.iis.sinica.edu.tw:9187))
- 實體辨識系統 ([deep.iis.sinica.edu.tw:9001](http://deep.iis.sinica.edu.tw:9001))
- 聊天機器人  
([learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/](http://learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/))
- 中文詞彙庫 ([ckip.iis.sinica.edu.tw:8080/license/](http://ckip.iis.sinica.edu.tw:8080/license/))
- ...

# E-HowNet瀏覽介面

廣義知網知識本體架構2.0版線上瀏覽

[授權辦法] 相關系統: 中文詞彙特性速描系統 | 斷詞系統 | 剖析系統

Collapse WIKI 說明

查節點 查定義式 查詢 進階:

查詢詞彙  
 查詢 taxonomy

Word (4)

- 蜻蜓
- 響蜻蜓
- 竹蜻蜓
- 蜻蜓點水

Taxonomy (1)

Category (0)

TopNode

- entity|事物
  - event|事件
  - object|物體 [ 事物, 客體, 對象 ]
    - thing|萬物 [ 天地萬物, 東西, 東東, 物類, 庶物, 萬有, 萬物, 群生 ]
      - physical|物質 [ 物產, 物質, 物體, 實體 ]
        - animate|生物 [ 生物, 生物體, 有機體, 物種, 活體, 動物體, 動植物 ]
          - AnimalHuman|動物
            - human|人 [ 人, 人氏, 人兒, 人物, 人類, 方趾圓顛, 者, 萬物之靈, 儕, 徒, 子 ]
            - animal|獸 [ 飛禽走獸, 動物, 野獸, 鳥獸, 禽獸, 鱗毛 ]
              - beast|走獸 [ 走獸, 獸, 獸類 ]
              - livestock|牲畜 [ 牲口, 牲畜, 家畜, 畜, 畜生, 畜牲, 畜類 ]
              - bird|禽 [ 鳥, 鳥兒, 鳥禽, 鳥類, 禽, 禽鳥, 禽類 ]
              - InsectWorm|蟲 [ 虫, 虫儿, 虫兒, 豸, 昆蟲, 蟲, 蟲子, 蟲兒, 蟲類 ]
                - 蠕蟲|worm [ 蠕蟲 ]
                - 甲蟲|beetle [ 甲蟲, 甲蟲類, 锹形蟲, 蟻 ]
                  - 蜻蜓|dragonfly [ 豆娘, 蜻蛉, 蜻蜓, 蜻蜓, 螞螂 ]
                  - 螢火蟲|firefly [ 流螢, 螢火蟲, 螢火蟲兒 ]
                - 蜂|bee [ 赤眼蜂, 金小蜂, 食子蜂, 馬蜂, 黃蜂, 蜂 ]
                - 蝴蝶|butterfly [ 天蝶, 白粉蝶, 海蝶, 蛺蝶, 鳳蝶, ]
                - 蛾|moth [ 石蠶, 飛蛾, 蛾, 蛾子, 蝶蠶, 燈蛾, 蠶蛾 ]
                - 蟬|cicada [ 知了, 秋蟬, 蚱蟬, 蜩, 螞蟬, 蜩, 蟬 ]
                - 竹節蟲|walkingstick [ 竹節蟲 ]
                - 蚱蜢|grasshopper [ 紡織娘, 蚱蜢, 蚱, 螞蚱, 蠡斯 ]
                - 蜘蛛|spider [ 蛛, 蛛蛛, 蜘蛛, 蜘蛛類, 蠅虎 ]
                - 蟋蟀|cricket [ 灶雞, 金鈴子, 蚰, 蚰蚰兒, 蛩, 鈴蟲 ]
                - 蠍子|scorpion [ 蝎, 蠍, 蠍子, 蠶 ]
                - 螳螂|mantis [ 螳, 螳螂 ]

詞彙訊息

|               |  |
|---------------|--|
| 詞彙:           | 蜻蜓   |
| 詞性:           | Nab  |
| 英文意涵:         | dragonfly  |
| Event Frame:  |  |
| 定義式:          | {蜻蜓 dragonfly}   |
| 操作式:          |  |
| 語義功能:         |  |
| 語義特徵:         | +worms&insects   |
| 展開式:          | {InsectWorm 蟲:predication={fly 飛;theme={~}},predication={eat 吃:patient={InsectWorm 蟲},agent={~}} |
| WordNet 自動連結: | {dragonfly.n.01}   |

# E-HowNet情緒分類

廣義知網知識本體架構2.0版線上瀏覽 (本介面僅呈現約三分之一的詞彙數量供參考查詢)

The screenshot displays the E-HowNet online browser interface. On the left, there is a navigation panel with a search bar and buttons for 'Collapse', '說明', '查節點', '查定義式', and '進階定義式查詢'. The main area shows a hierarchical tree structure starting from 'TopNode'. The tree is expanded to show the following categories and their associated terms:

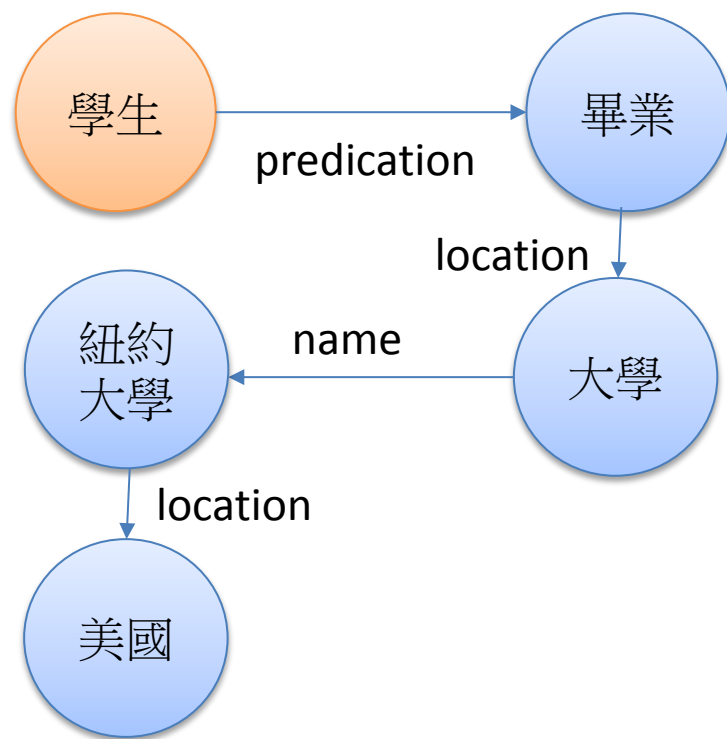
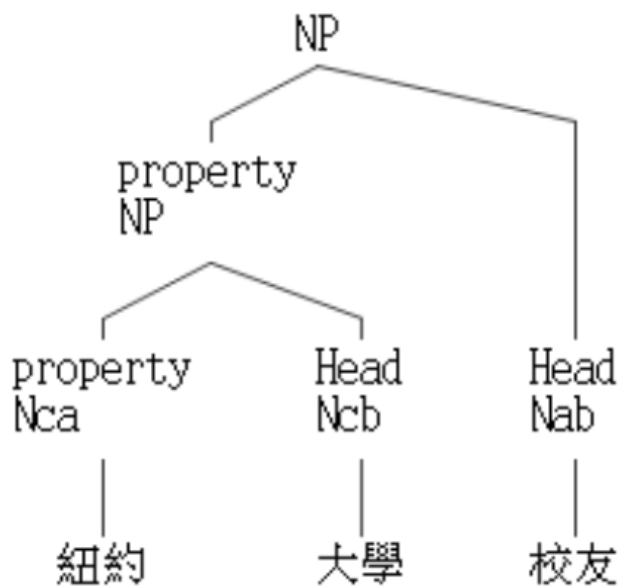
- entity|事物
  - event|事件
    - state|狀態
      - PhysicalState|物理狀態
      - MentalState|精神狀態
        - FeelingByGood|好心情
          - AtEase|安心 [ 安心, 放心, 安慰, 不虞, 釋懷, 心安 ]
          - joyful|喜悅 [ 高興, 樂, 快樂, 喜, 歡, 歡樂, 開心, 喜悅, 欣, 歡喜, 驚喜, 欣然, 欣喜, ]
          - satisfied|滿意 [ 滿意, 滿足, 欣慰, 如意, 中意, 自足, 合意, 自得, 心滿意足, 安於, 知足 ]
          - FeelNoQualms|無愧 [ 問心無愧, 理直氣壯, 無愧 ]
          - shameless|沒羞 [ 無恥, 好意思, 不要臉, 厚顏, 厚臉皮, 寡廉鮮恥 ]
        - FeelingByBad|壞心情
          - uneasy|不安 [ 不安, 不知所措, 失色, 風聲鶴唳, 懸疑, 忐忑不安, 打鼓, 侷促, 惴惴不安, 坐 ]
          - unsatisfied|不滿 [ 不滿, 不服, 適得其反, 心生不滿, 事與願違 ]
          - upset|煩惱 [ 頭痛, 煩惱, 煩, 苦惱, 愁, 想不開, 抑鬱, 頭大, 苦悶, 鬱卒, 操心, 煩躁, 惱 ]
          - sad|憂愁 [ 慮, 慮, 憂心忡忡, 憂鬱, 鬱悶, 傷感, 氣結, 憂懼, 自傷, 憂愁, 感感, 惆悵 ]
          - sorrowful|悲哀 [ 沈重, 難過, 痛心, 沉重, 酸, 創傷, 悲傷, 悲憤, 傷心, 悲痛, 悲, 悲哀, ]
          - fear|害怕 [ 恐, 怕, 恐懼, 恐慌, 害怕, 驚, 怖, 驚慌, 畏, 唯恐, 畏懼, 深怕, 虛驚, 懼, , 震懾, 心驚, 惘, 聞之色變, 驚心, 生怕, 驚懼, 心驚膽跳 ]
          - surprise|驚奇 [ 驚人, 驚訝, 訝異, 驚嚇, 錯愕, 驚奇, 爆炸性, 吃驚, 咋舌 ]
          - worried|著急 [ 擔心, 憂慮, 憂心, 掛, 擔憂, 顧慮, 憂, 焦慮, 焦, 後顧之憂, 焦急, 耽心, ]
          - angry|生氣 [ 氣, 火, 憤怒, 生氣, 氣憤, 憤慨, 怒, 變色, 震怒, 發飆, 忿怒, 出氣, 怒火, 瞋 ]
          - disheartened|灰心 [ 不振, 低落, 低潮, 氣餒, 涼, 灰心, 洩氣, 頹廢, 心灰意冷, 灰頭土臉, ]
          - repent|懊悔 [ 後悔, 恨, 憾, 悔, 懺悔, 懊惱, 懊悔, 悔悟 ]
          - shy|羞愧 [ 內向, 恥, 害羞, 腼腆, 羞, 羞澀, 面紅耳赤, 羞怯, 嬌羞 ]
          - embarrassed|為難 [ 尷尬, 困窘, 難堪, 不好意思, 拘束, 啼笑皆非, 窘狀, 哭笑不得 ]
          - disappointed|失望 [ 失望, 沮喪, 向隅, 扼腕, 失意, 心寒 ]
          - 慚愧|ashamed [ 恥辱, 慚愧, 羞恥, 愧對, 愧疚, 汗顏, 羞愧, 愧, 見不得人, 有愧, 為恥 ]

# Reference on E-HowNet

- Keh-Jiann Chen, Shu-Ling Huang, Yueh-Yin Shih, Yi-Jun Chen, 2005, Extended-HowNet- A Representational Framework for Concepts, Ontologies and Lexical Resources IJCNLP-05 Workshop
- Shu-Ling Huang, You-Shan Chung, Keh-Jiann Chen. , 2008, E-HowNet: the Expansion of HowNet, The First National HowNet Workshop
- Su-Chu Lin, Shu-Ling Huang, You-Shan Chung and Keh-Jiann Chen, 2013, The Lexical Knowledge and semantic representation of E-HowNet, Contemporary Linguistics, Vol.15, No.2, pp. 177-194.
- Yueh-Yin Shih, Wei-Yun Ma, Extended HowNet 2.0 – An Entity-Relation Common-Sense Representation Model, LREC 2018
- ...

# 目前研究

- 利用剖析結果自動化產生E-HowNet表達式
  - Ex: 紐約大學校友



謝謝聆聽



# Backup

# Introduction to Word Representation

# Word Representations

- Dictionary-based Word Representation  
(Discrete Word Representation)

|       |  |
|-------|--|
| 詞彙:   | 蝴蝶   |
| 詞性:   | Nab  |
| 英文意涵: | butterfly  |
| 概念式:  | {蝴蝶   butterfly}   |
| 展開式:  | {InsectWorm   蟲 : predication = {fly   飛 : theme = {~}}} |

# Is there other ways to represent a word?

In vector space terms, this is a vector with one 1 and a lot of zeroes

[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a “one-hot” representation. Its problem:

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND  
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

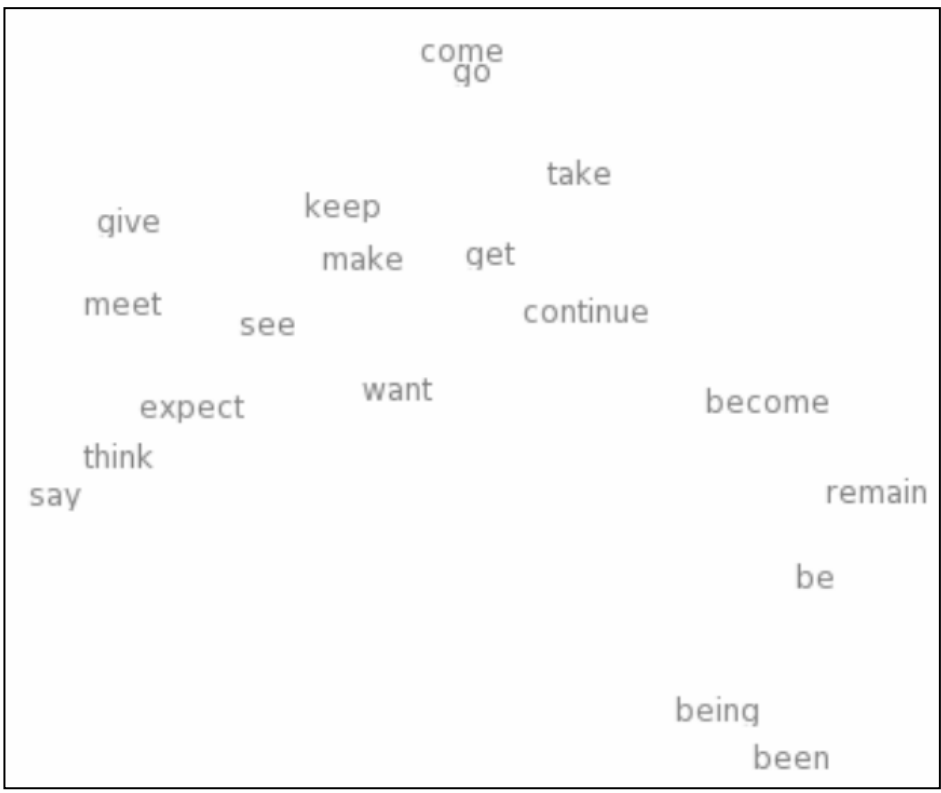
# Continuous Word Representations

- A word is represented as a dense vector

$$\text{蝴蝶} = \begin{bmatrix} 0.234 \\ 0.283 \\ -0.435 \\ 0.485 \\ -0.934 \\ -0.384 \\ 0.234 \\ 0.548 \\ -0.834 \\ 0.437 \\ 0.483 \end{bmatrix}$$

# Continuous Word Representations

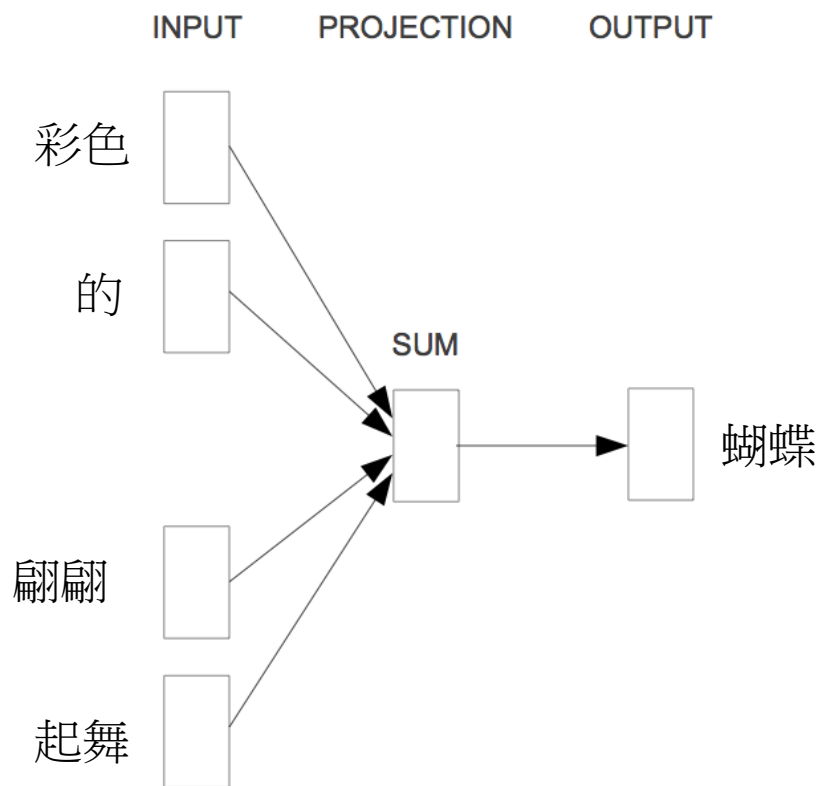
- Word embedding captures the word meaning and project the meaning into a semantic vector space



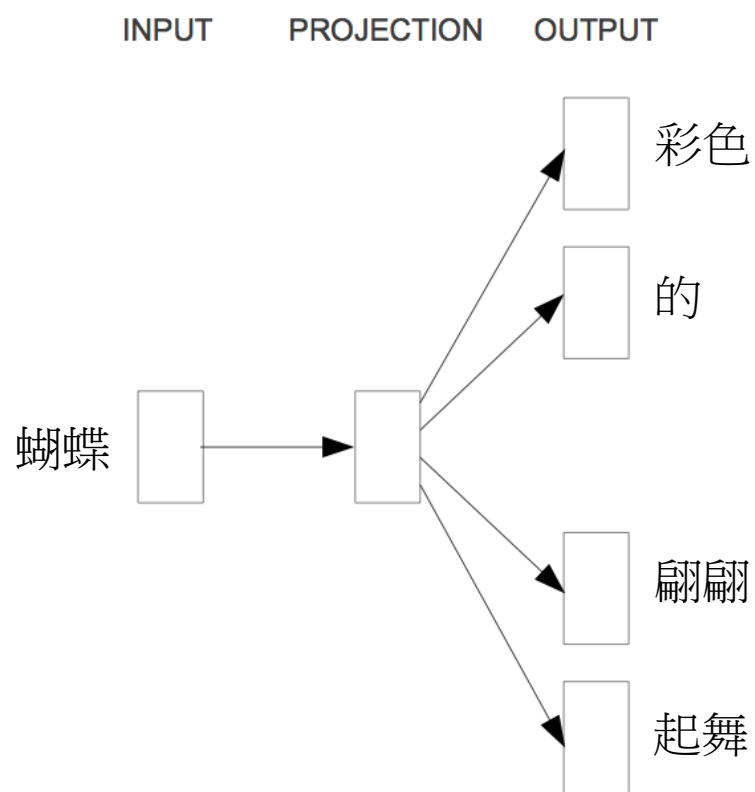
Adapted picture from "Deep Learning for NLP(without Magic)" of Socher and Manning

# Word2vec

彩色 的 蝴蝶 翩翩 起舞



**Continuous Bag-of-Words (CBOW)**



**Skip-gram**

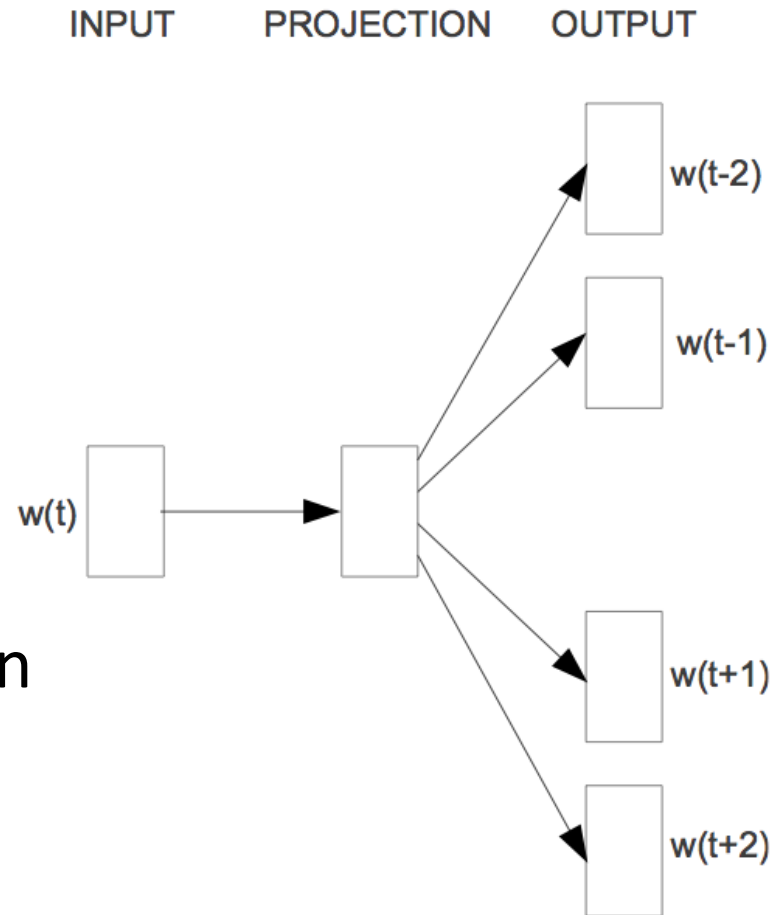
# Skip-gram using softmax

- The objective of the Skip-gram model is to maximize the average log probability

$$\frac{1}{T} \sum_{t=1}^T \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$$

- The basic Skip-gram formulation defines  $p(w_{t+j} | w_t)$  using the softmax function:

$$p(w_O | w_I) = \frac{\exp(v'_{w_O} \top v_{w_I})}{\sum_{w=1}^W \exp(v'_w \top v_{w_I})}$$



The probability is also used on training. But so many calculations!



# Skip-gram using Negative sampling

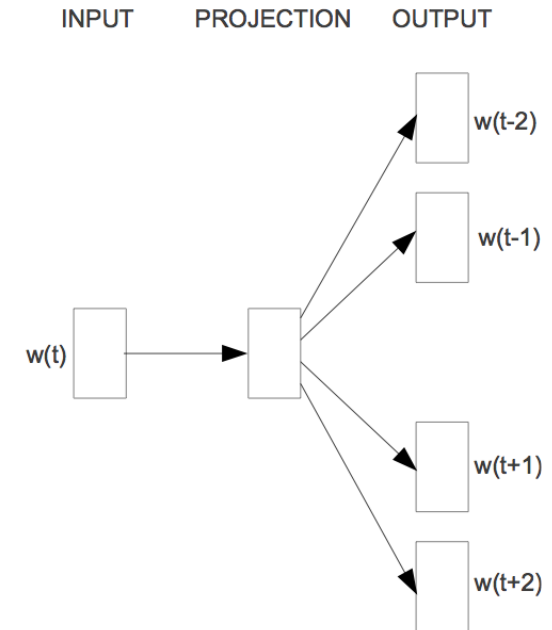
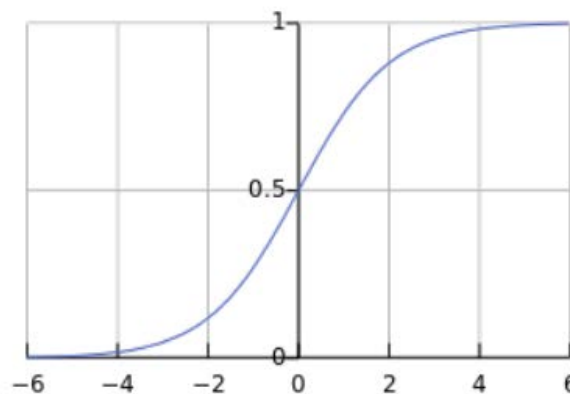
- We define Negative sampling (NEG) by the objective

$$\log \sigma(v'_{w_O} \top v_{w_I}) + \sum_{i=1}^k \mathbb{E}_{w_i \sim P_n(w)} \left[ \log \sigma(-v'_{w_i} \top v_{w_I}) \right]$$

k is the number of negative samples

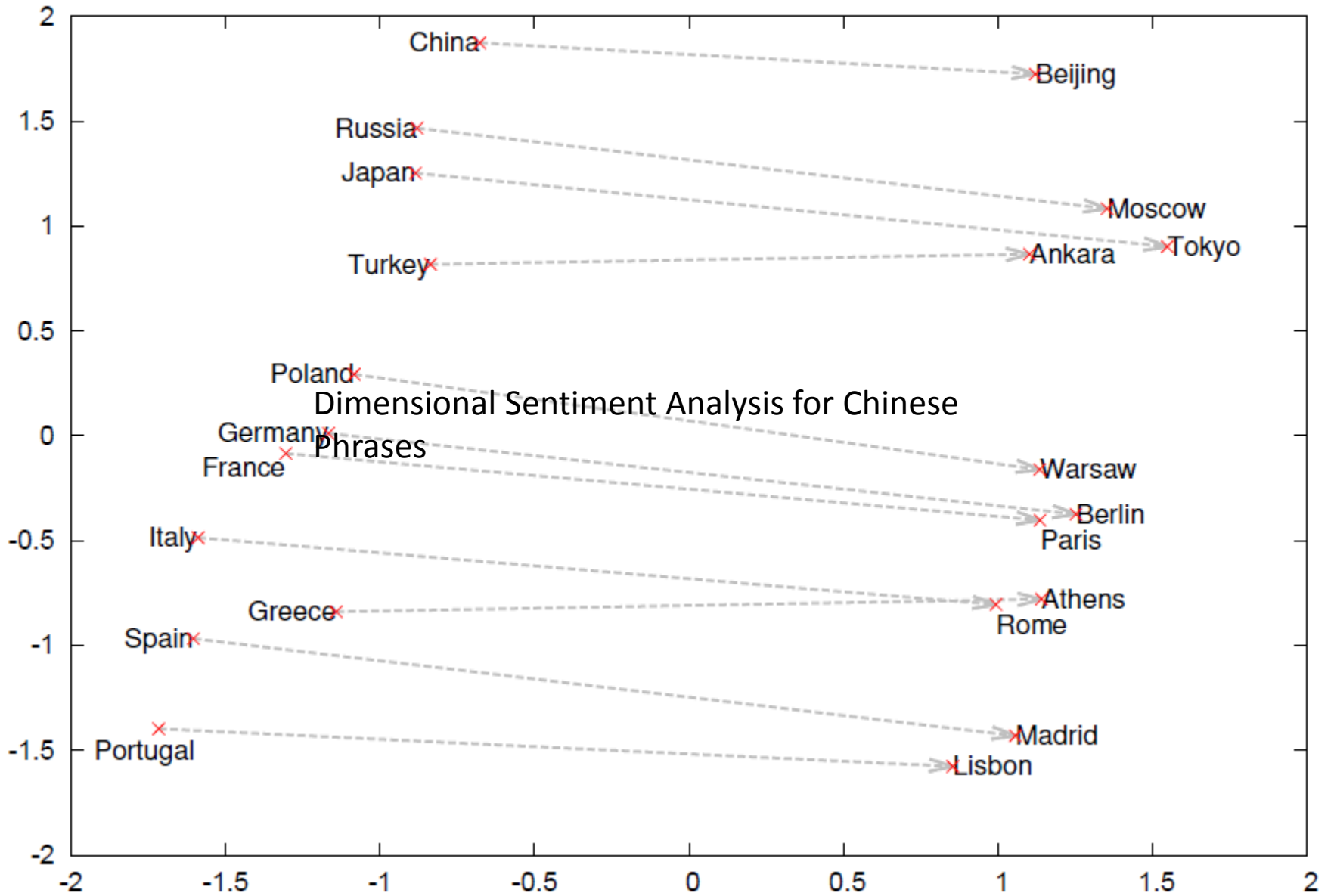
- The sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



← The function is also used on training.  
So little calculations!

# Country and Capital Vectors Projected by PCA



(Mikolov et.al, NIPS 2013)

# Introduction to Chatbot

# Chatbot的功能

- Task Completion
- Social Chat (chit-chat)
- Negotiation
- Game
- ...

# Task Completion

- 完成開放領域的任務, Ex: Google查詢



Google

美國總統是誰?

全部 新聞 圖片 影片 地圖 更多 設定 工具

約有 15,200,000 項結果 (搜尋時間：0.63 秒)

美國 / 總統

唐納·川普

更多唐納·川普的相關資訊

# Task Completion

- 完成特定領域的任務, Ex: QA查詢, 訂票

The screenshot shows a chat interface for CTBC Bank. At the top, there is a green header with the bank's logo and name. Below the header, a light blue bar contains the text '個資宣告' (Privacy Statement) and a checkmark icon. The chat history consists of several messages:

- User:** 我想知道現在利率? (I want to know the current interest rate?) - 08:33:32
- 小C:** 利率查詢方式區分如下, 請點選您要了解的內容:  
1. 銀行帳戶  
2. 信用卡  
3. 信貸  
4. 房貸 - 08:33:34
- User:** 銀行幾點開門 (When does the bank open?) - 08:34:28
- 小C:** 全省分行營業時間: 營業日上午09:00 ~ 15:30, 目前無假日分行服務。  
小C提供您以下相關問題參考:  
1. [數位帳戶申請方式](#) - 08:34:30
- User:** 你叫什麼名字 (What is your name?) - 08:35:00
- 小C:** 小C, 因為我每天都笑嘻嘻, 希望你也能每天心情愉快喔! - 08:35:02

中研院怎麼去



08:35:14



08:35:16

哎呀呀~小C不太理解您的意思，建議您可以重新輸入20字以內的問題詢問小C，例如:繳款金額是多少?小C就能幫忙您唷~  
~

我覺得你們利率太低了



08:45:26



08:45:28

請您輸入20個字以內的單一問句描述問題，或是參考以下問題：

1. [貸款專案說明](#)
2. [各項利率查詢](#)
3. [存款利率查詢](#)

\* 請輸入單一數字了解詳細內容

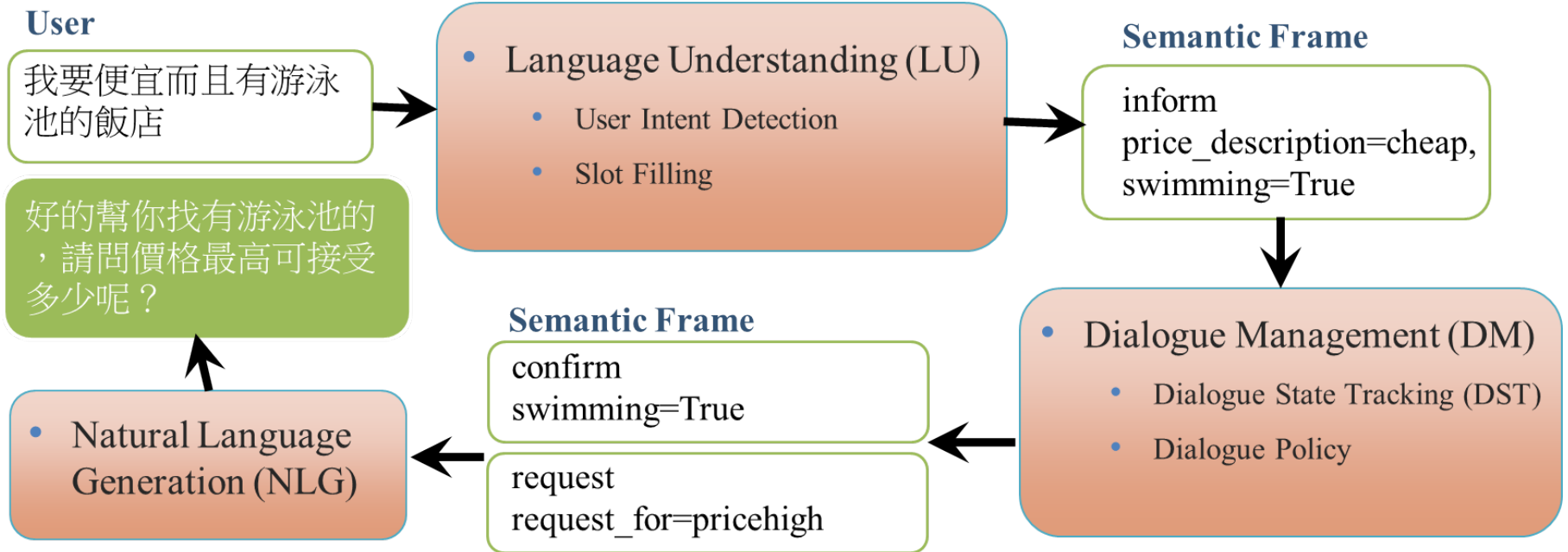
# Social Bot

- 聊天 Ex: 微軟小冰





# Chatbot的架構設計



# 建立商品屬性表

- 決定slot以及slot value or slot value type, 並爬文建構. 以書為例：

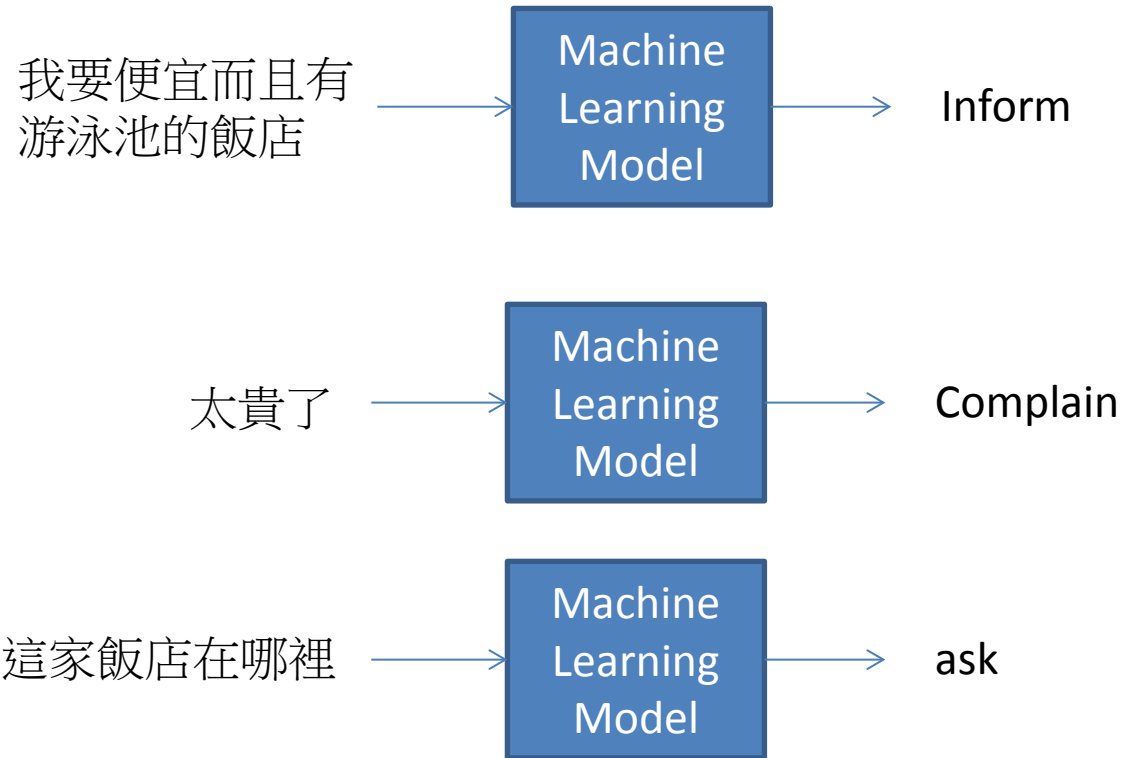
|       | 價錢    | 商店   |
|-------|-------|------|
| 射鵬英雄傳 | \$840 | 博客來  |
| 射鵬英雄傳 | \$870 | MOMO |

動態商品屬性表

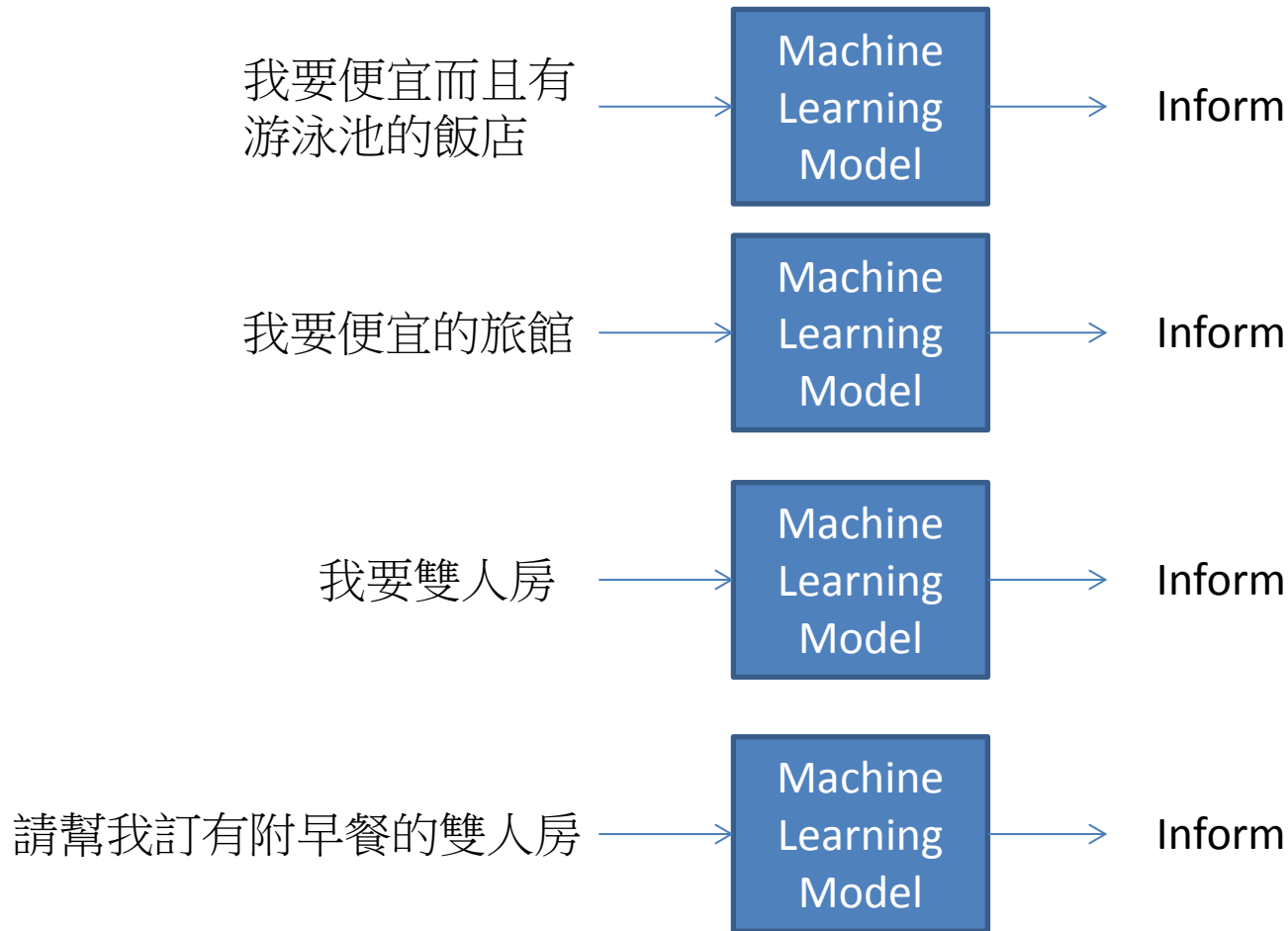
|       | 作者          | 出版社 | 簡介  | 書中人物  | 翻譯作品 | 出版日期 |
|-------|-------------|-----|---|-------|------|------|
| 射鵬英雄傳 | 金庸          | 遠流  | 《射鵬》最成功之處，是在人物的創造。《射鵬》的故事，甚至可以說是平鋪直敘的，所有精采的部分，全來自所創造出來的、活龍活現、無時無刻不在讀者眼前跳躍的人物，如眾所周知的郭靖、黃蓉。 | 郭靖、黃蓉 | 否    | 1957 |
| 冷山    | Frazier, C. | 書林  | 美國南北戰爭末期的一名受傷士兵英曼，在靈魂彷彿燒盡之後，獨行踽踽地回到家鄉；他的情人艾姐，則在山影交錯的鄉間度過了失怙獨立的蛻變期，學會如何與粗礪尖銳的生活對抗掙扎。       | 英曼、艾姐 | 是    | 1900 |

靜態商品屬性表

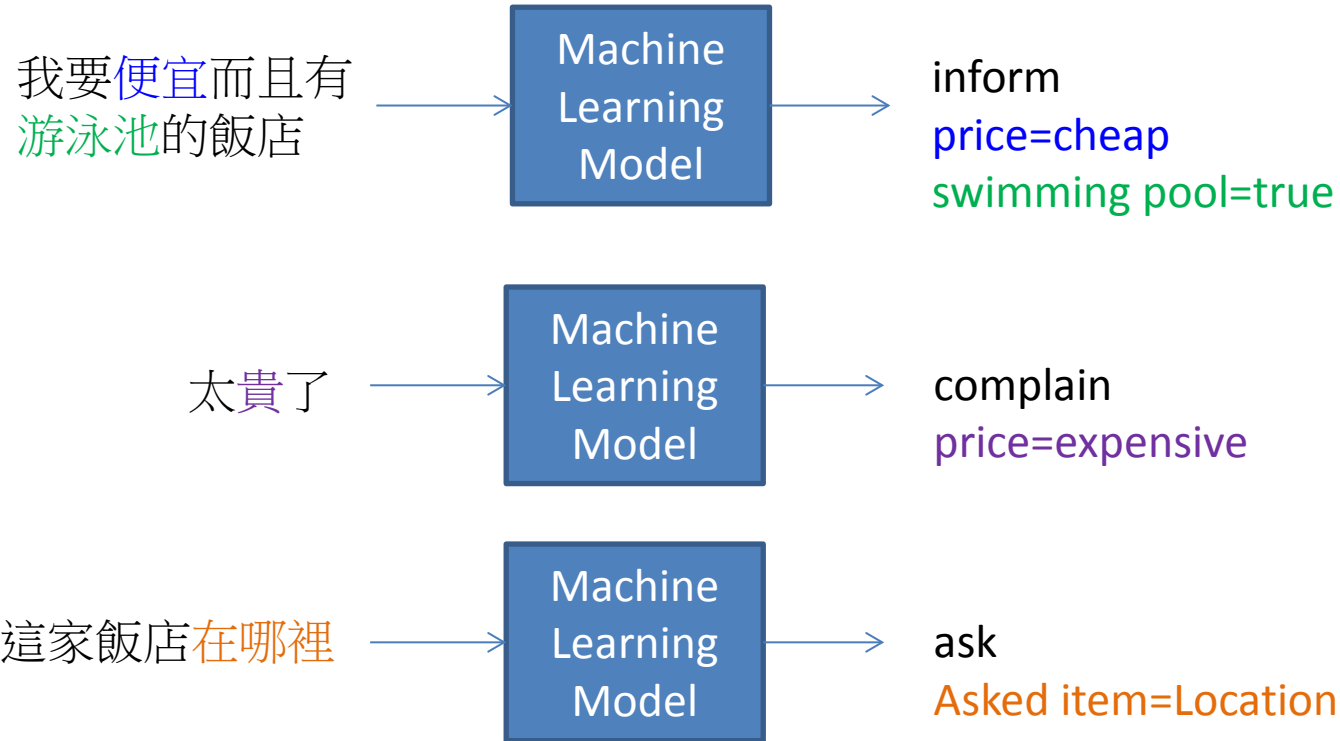
# Intention Prediction



# Training for Intention “Inform”

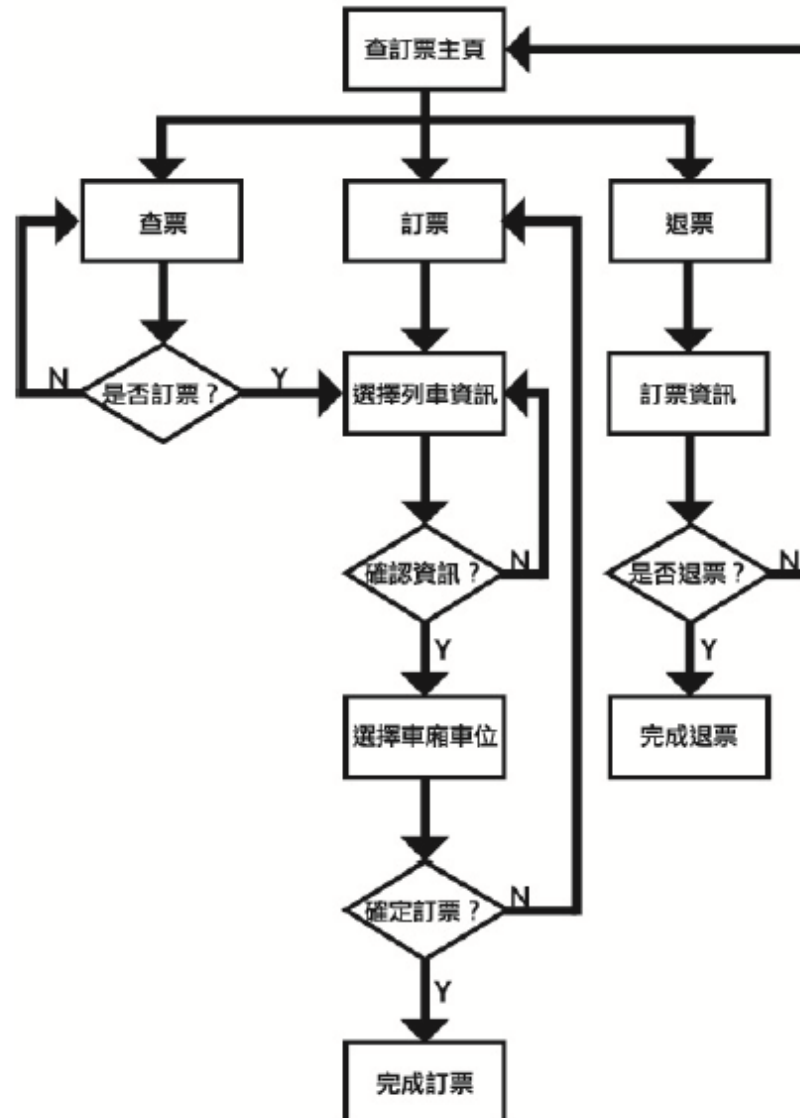


# Semantic frame: Intention + Slot Filling



# Dialog Management

台鐵訂票系統流程圖



# Chatbot的語言生成技術分類

- 罐頭文字(Canned Text)
- 基於模板的生成(Template-based Generation)
- 實例擷取的方法(Retrieval-based Approach)
- 根據語義表達式進行語言生成(Generation-based Approach via Semantic Frame)
- 字串編碼及解碼進行語言生成(Generation-based Approach via Sequence-to-Sequence Model)