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

2018/05/19

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

- 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)

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)

Chinese NLU by CKIP

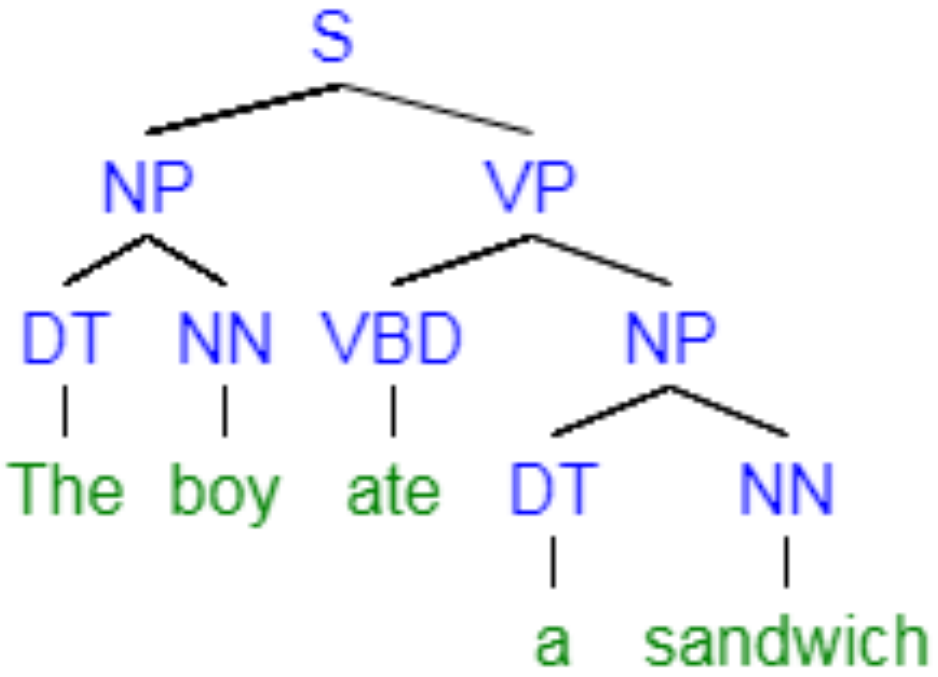
- 中文斷詞系統 **(ckipsvr.iis.sinica.edu.tw)**
- 中文剖析系統 **(parser.iis.sinica.edu.tw)**
- 中文詞彙特性速描系統 **(wordsketch.ling.sinica.edu.tw)**
- 廣義知網線上系統 **(ehownet.iis.sinica.edu.tw)**
- 輿情分析系統 **(learn.iis.sinica.edu.tw:9187)**
- 實體辨識系統 **(deep.iis.sinica.edu.tw:9001)**
- 聊天機器人 **(learn.iis.sinica.edu.tw/~dgrey1116/chatbot-demo/)**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

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)

Syntactic Structures by Syntactic Parsing

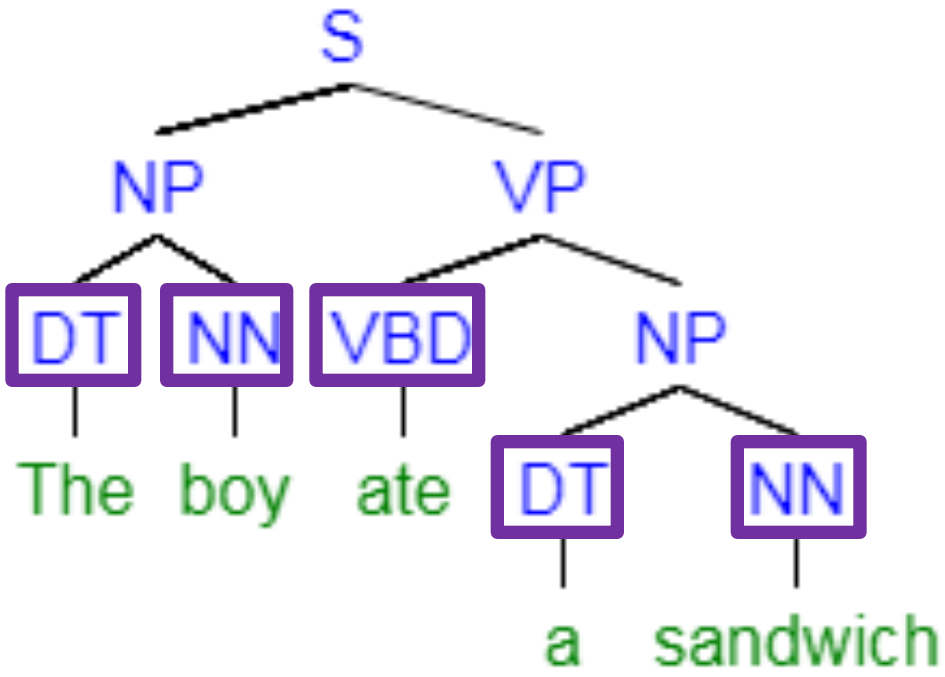
- Provide syntactic structure for a given sentence



The Information Conveyed by Parse Trees

(1) Part of speech for each word

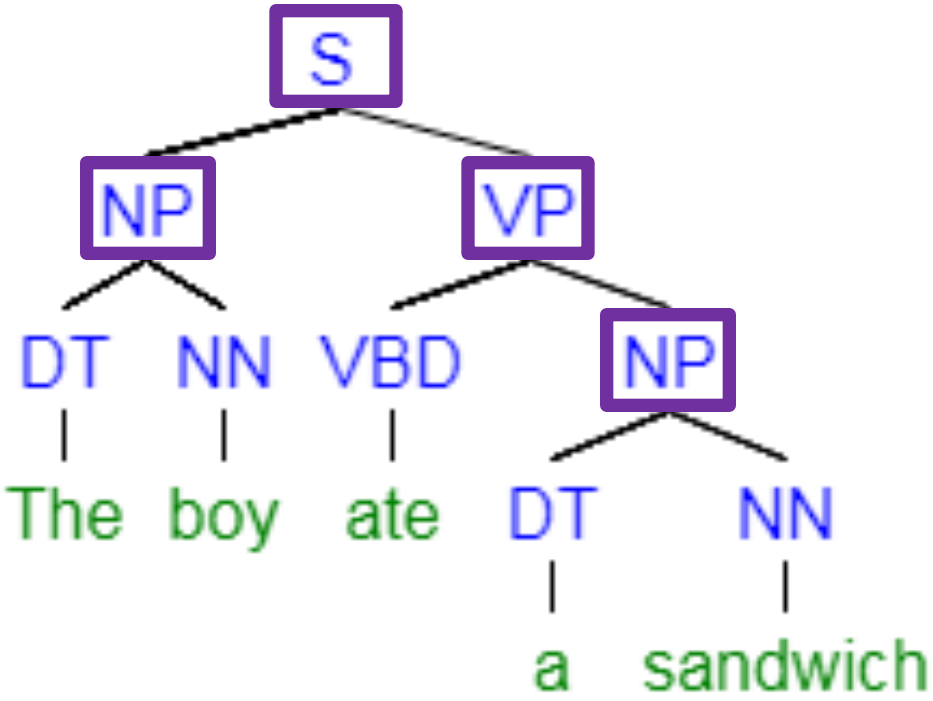
NN = noun, VBD = verb(past tense), DT = determiner)



The Information Conveyed by Parse Trees

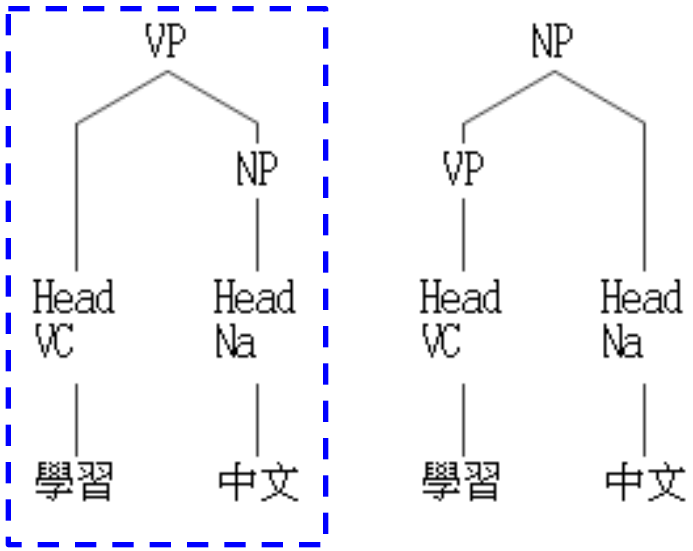
(2) Phrases

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



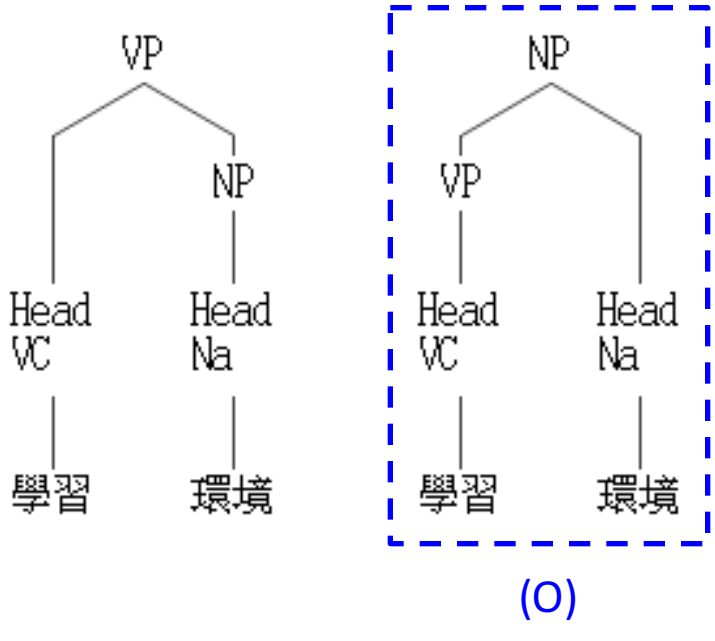
Syntactic Ambiguity

學習(Vt)　中文(N)



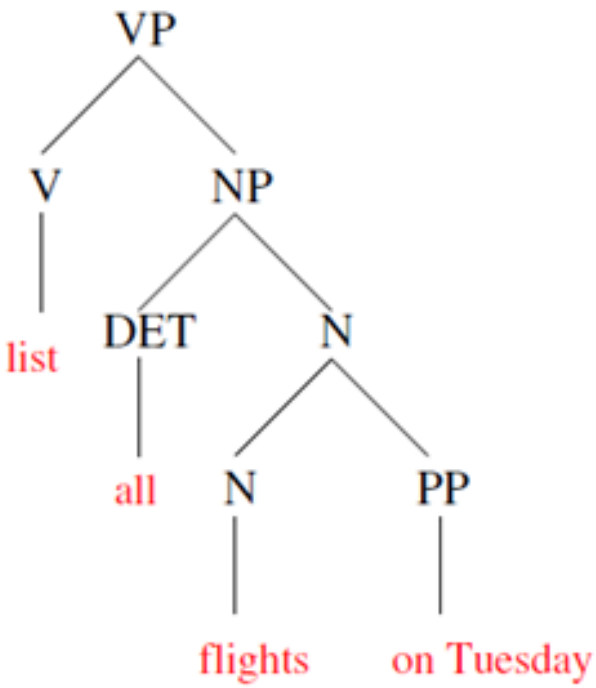
(O)

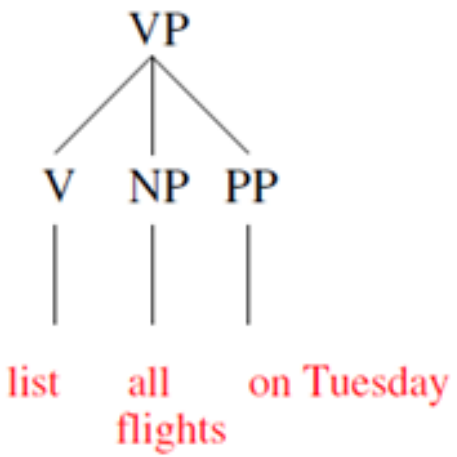
學習(Vt)　環境(N)



(O)

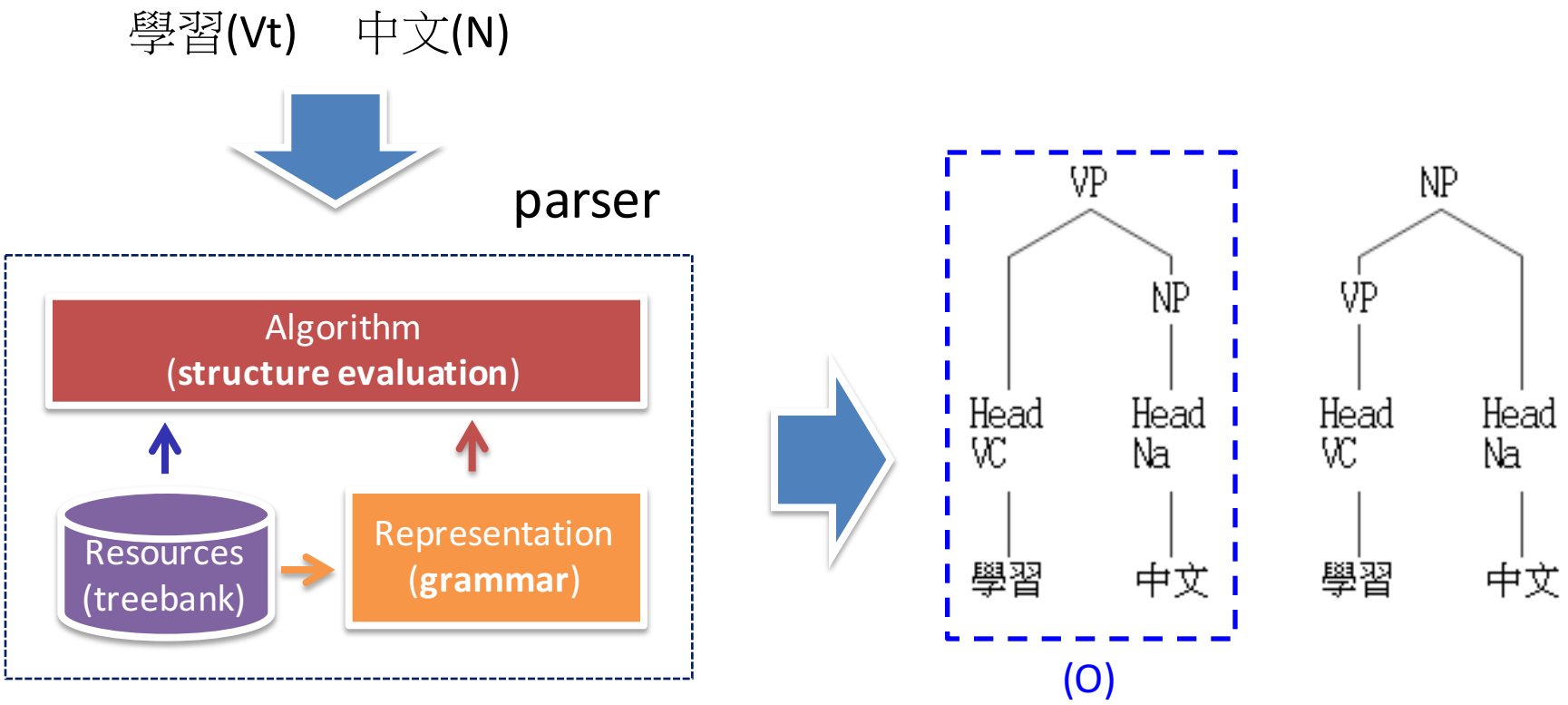
Syntactic Ambiguity





Example from Michael Collins

Parsing = Structure Disambiguation



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)

Outline

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

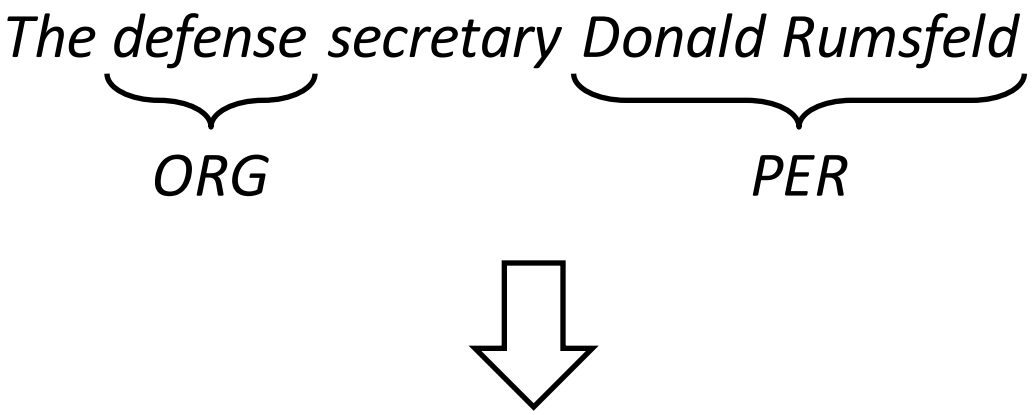
To Locate and Classify Named Entities (NEs)

The defense secretary Donald Rumsfeld

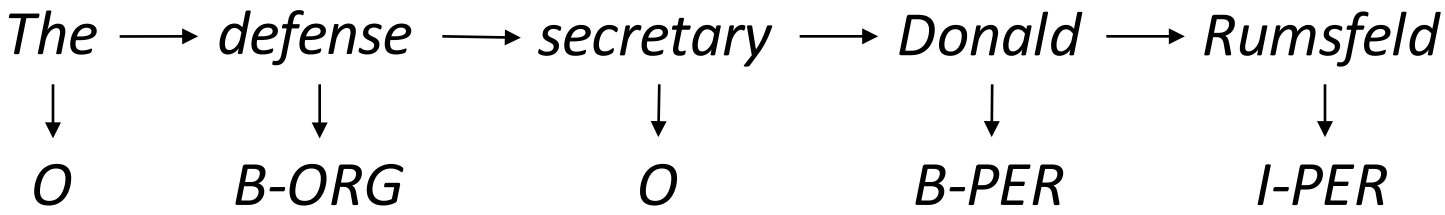
ORG *PER*

NE Labels
PER (person) *ORG (organization)*

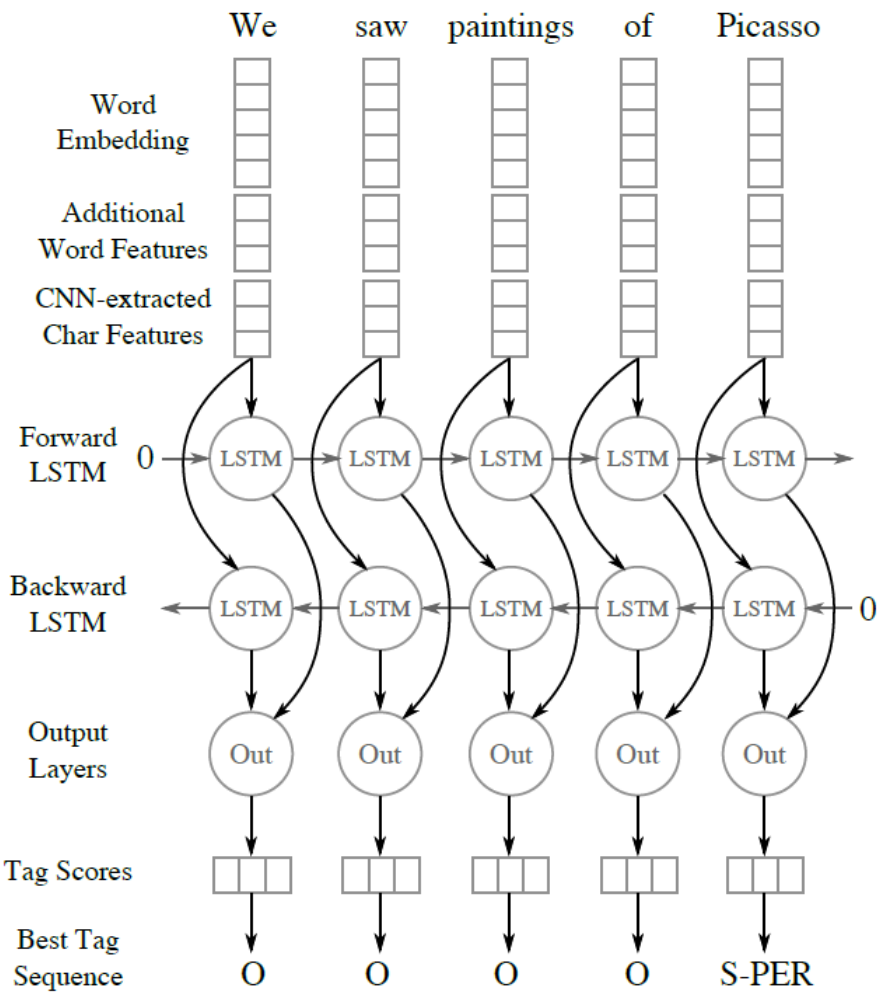
Popular Approaches: Sequential Labeling NER (Seq-NER)

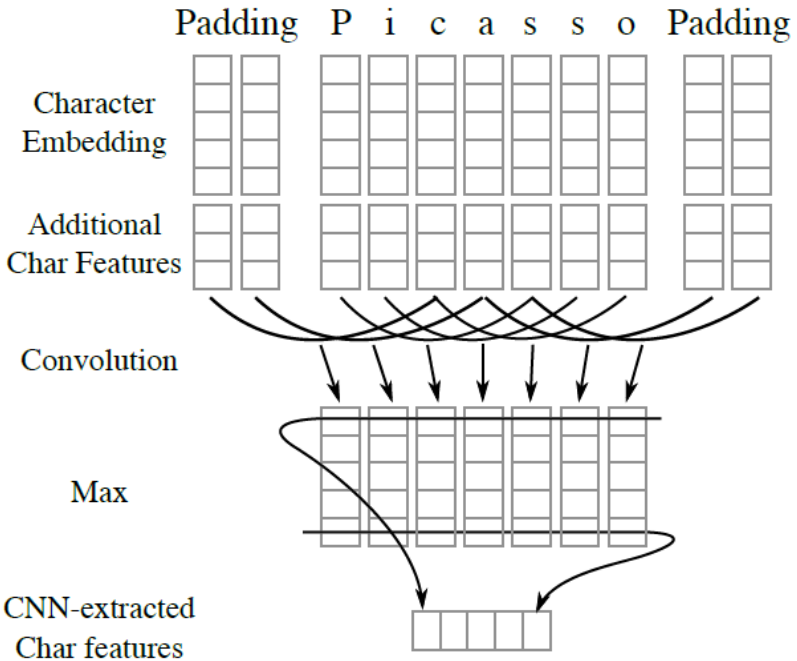


Seq Labels
B (begin)
I (inside)
O (outside)



Chiu and Nichols (2016)





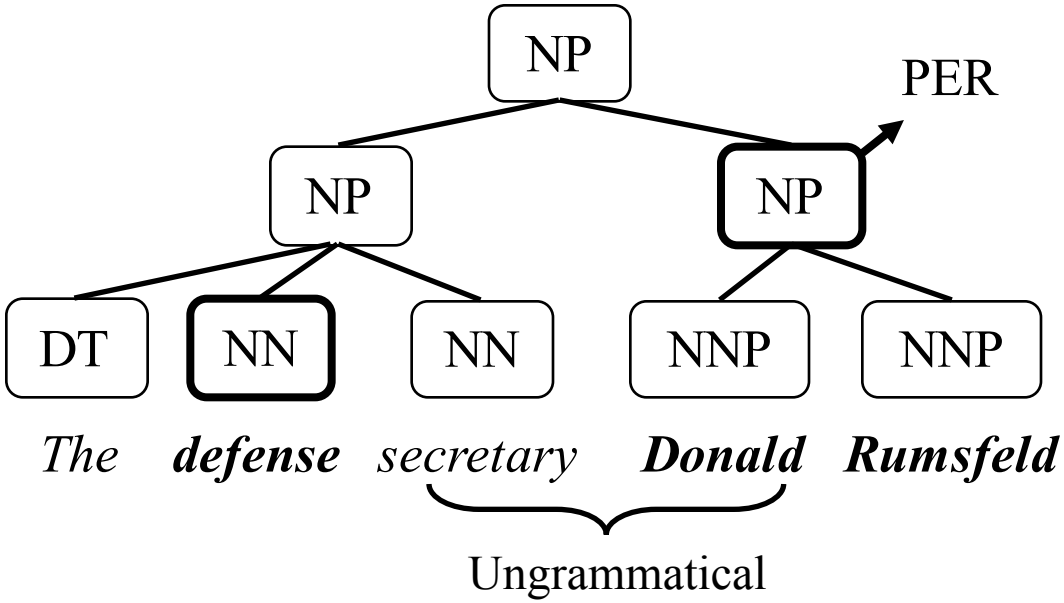
J. P. Chiu and E. Nichols. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

Outline

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

Motivation

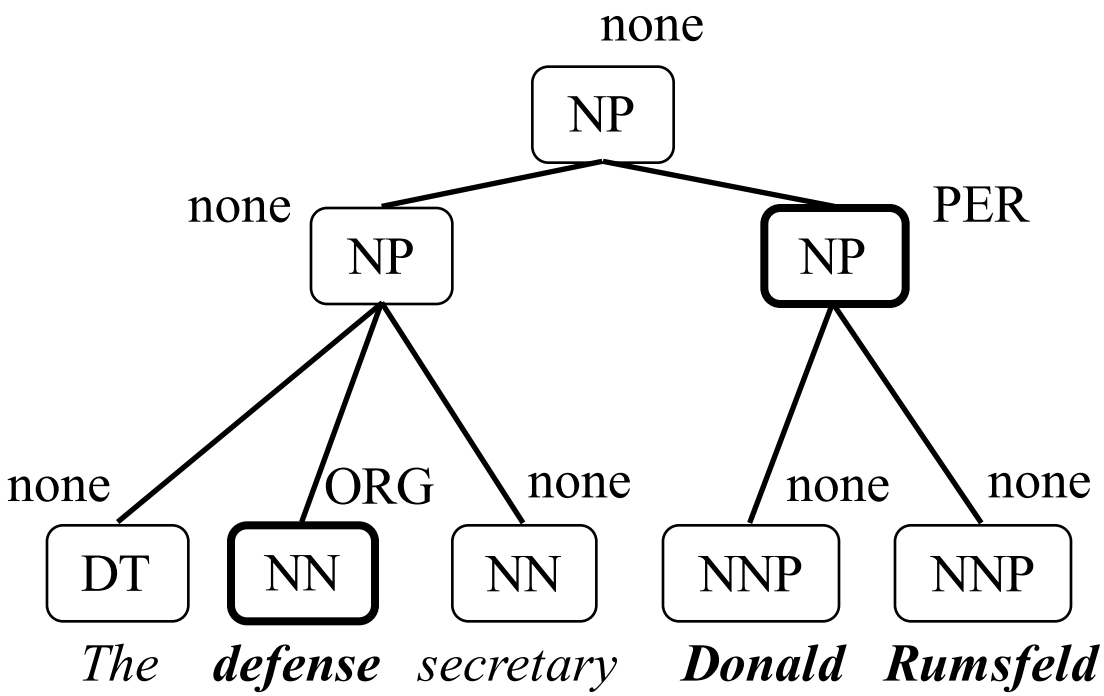
- Our observation
  - most NEs are constituents



Constituent → A plausible NE candidate

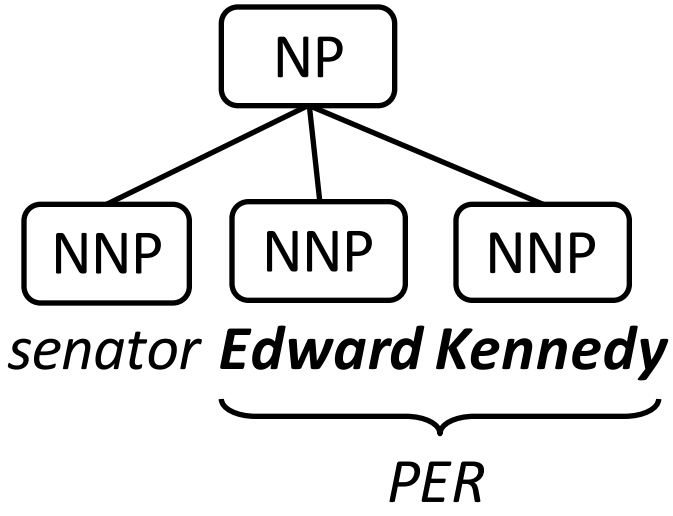
Ungrammatical → Unlikely an NE

Our Idea: NER through labeling each constituent

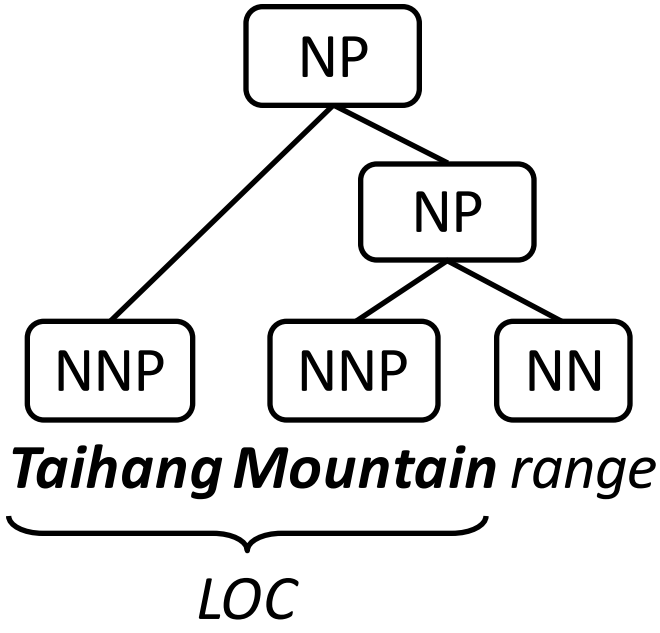


But…

- There are still some inconsistent NE

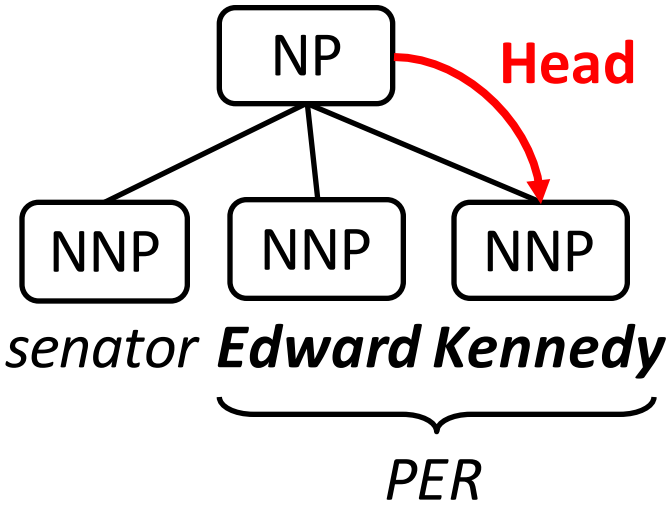


Type-1
Cross Siblings

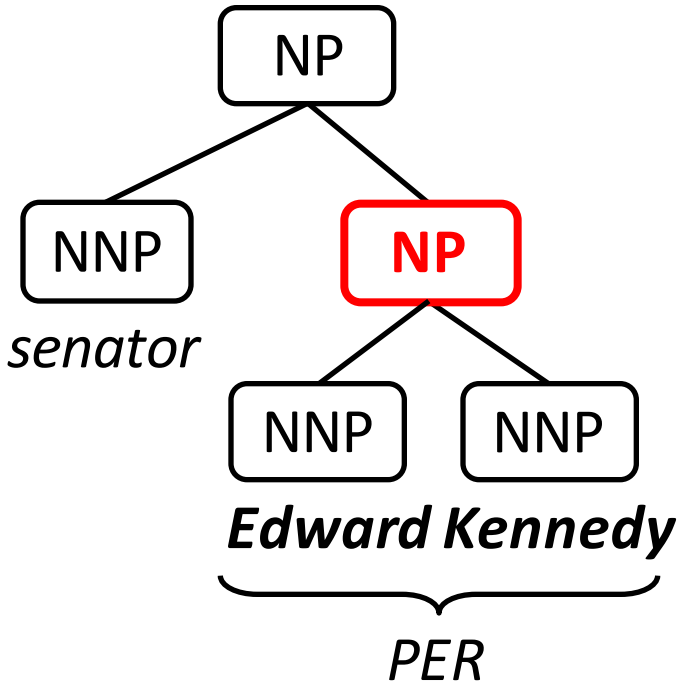


Type-2
Cross Branches

Eliminate Type-1: Constituency Tree Binarization

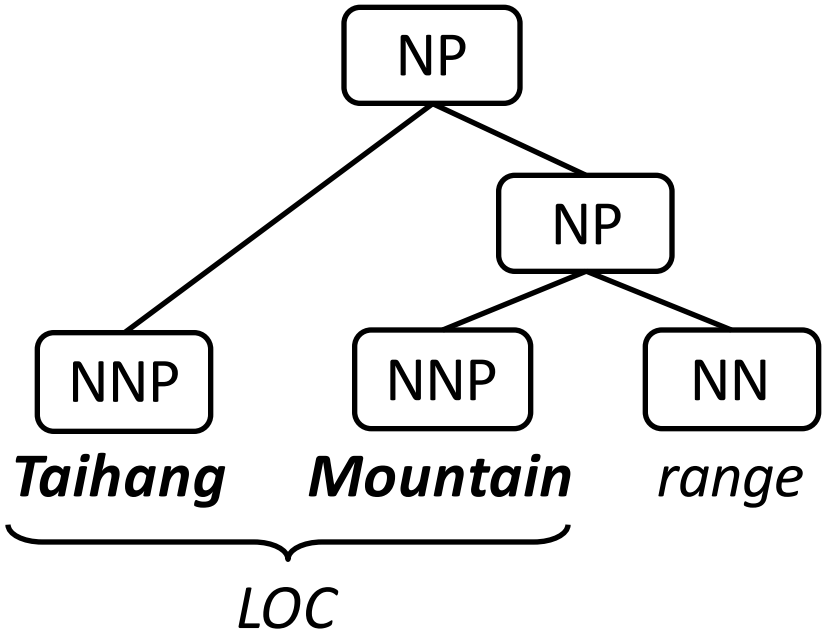


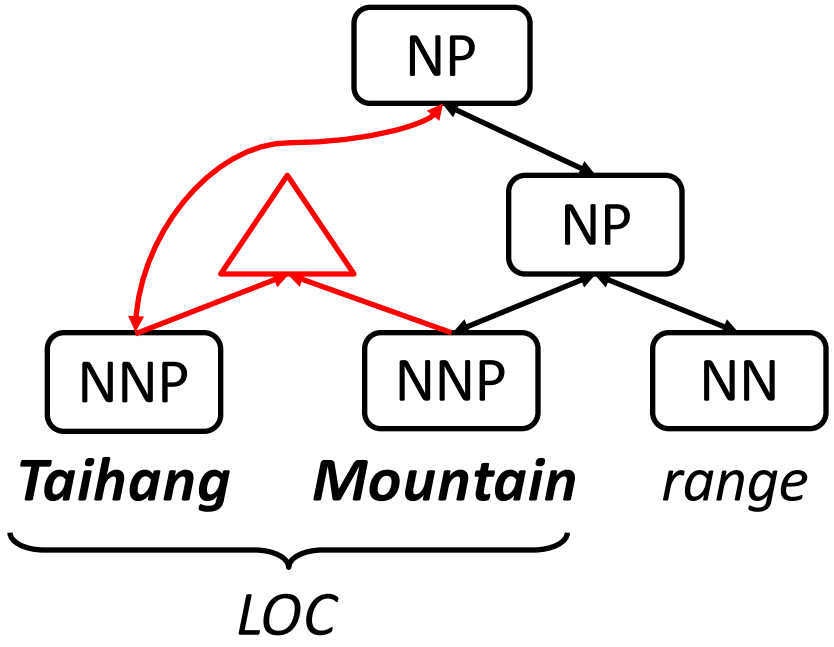
Type-1
Cross Siblings



Consistent

Eliminate Type-2: Pyramid Construction

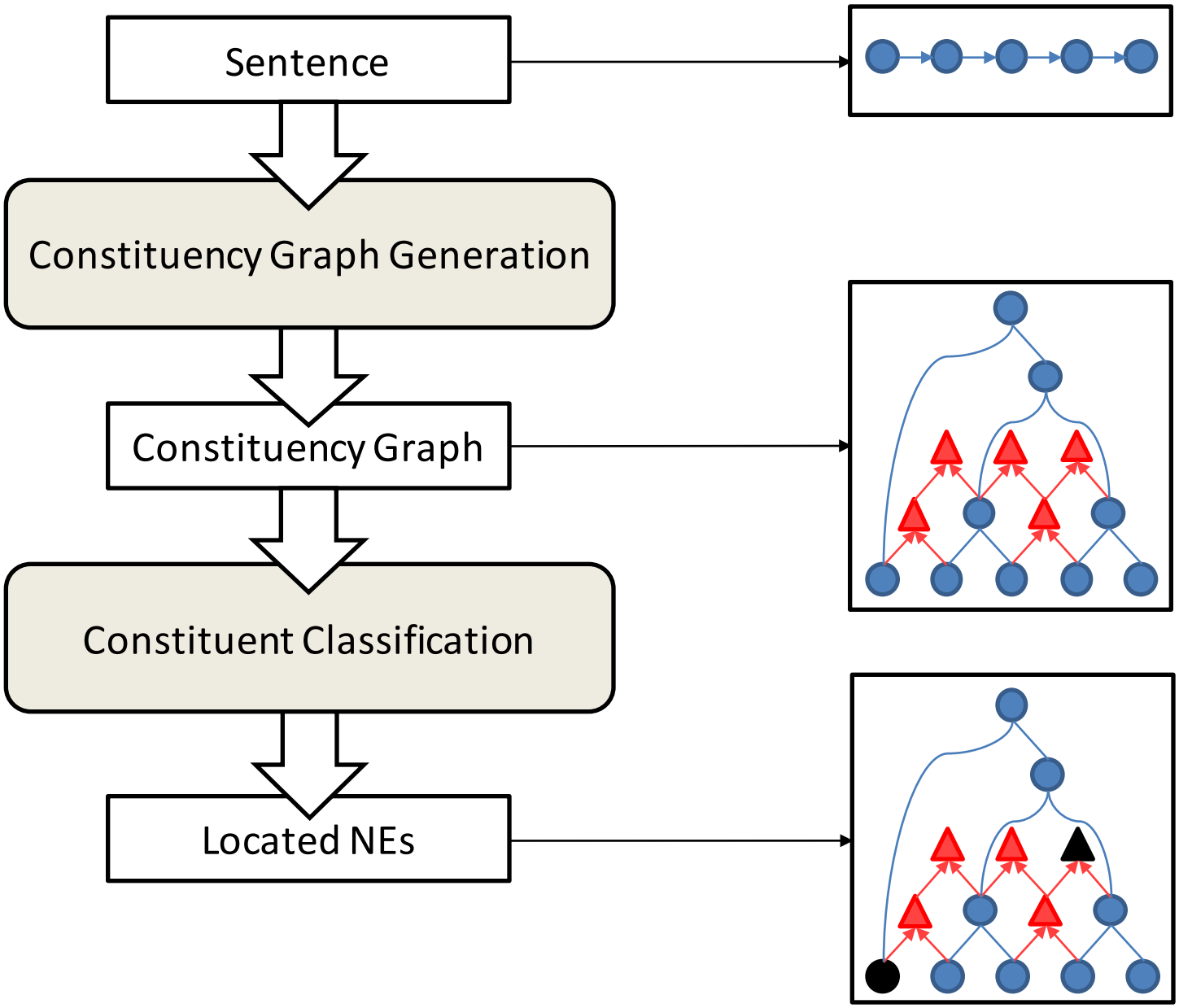




Type-2
Cross Branches

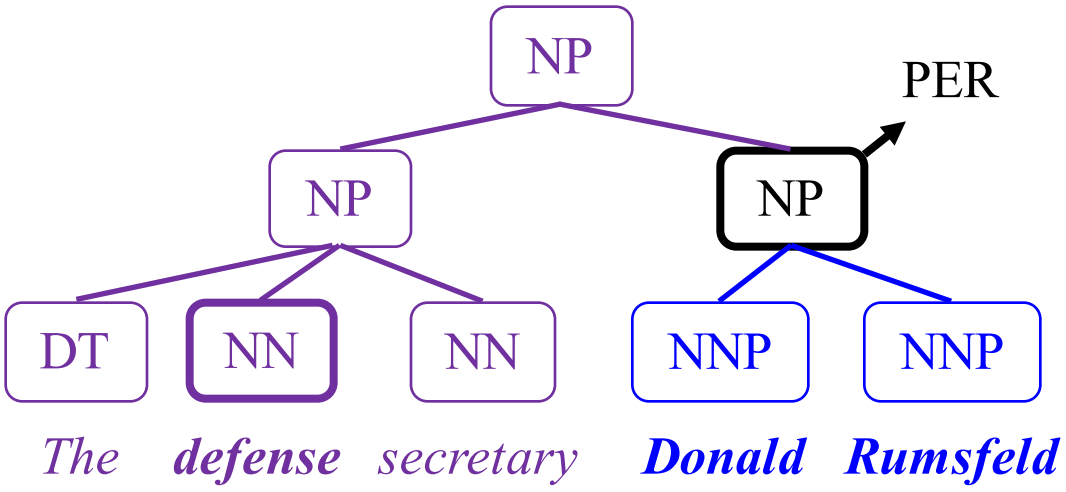
No Inconsistencies

Constituency-Oriented NER

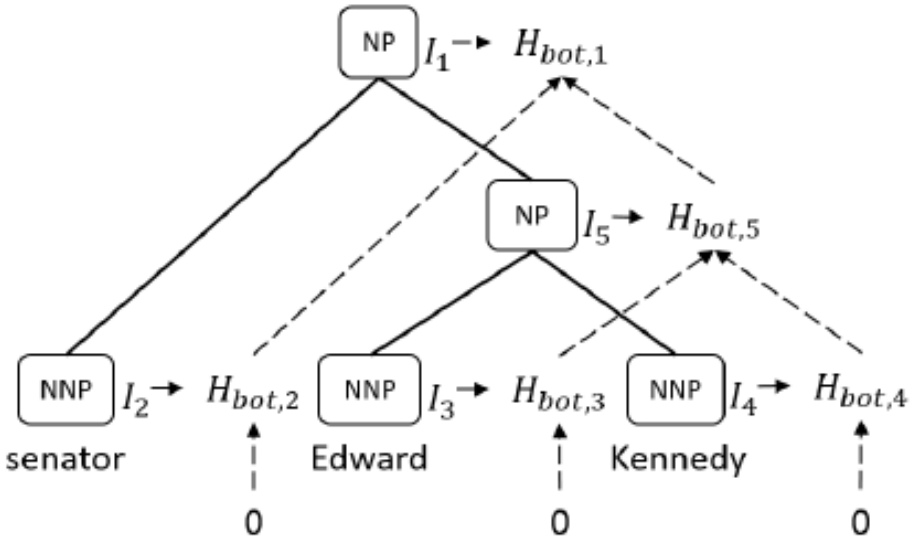


Constituent Classification

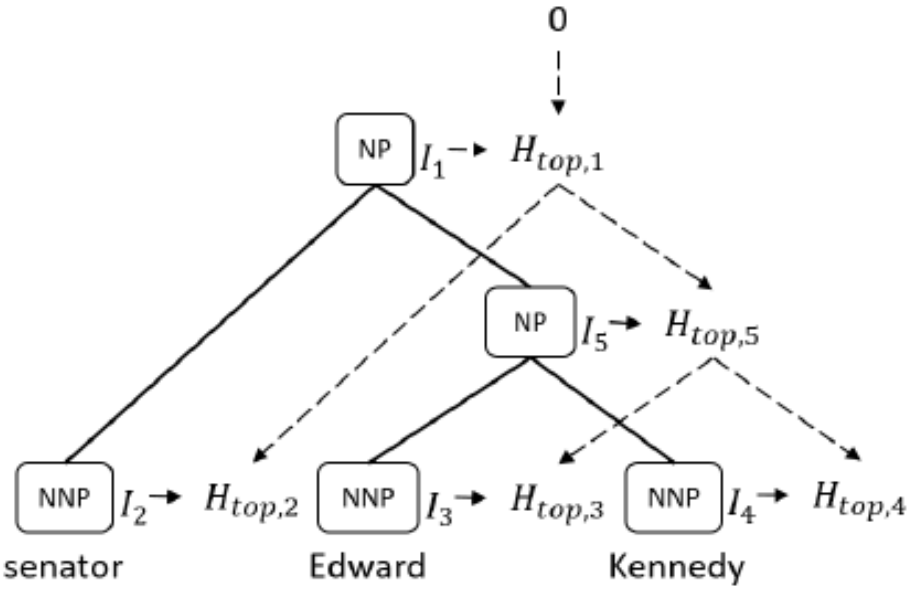
- Constituent Labelling should be using "inside" and "outside" information of the target constituent



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

$$O_i = Softmax((H_i \| H_l \| H_r) W_{out} + b_{out})$$

Seq-Recurrent vs. Constituency-Oriented BRNN

93% Consistency **97% Consistency**

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

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

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)

Demo

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

Demo

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

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)

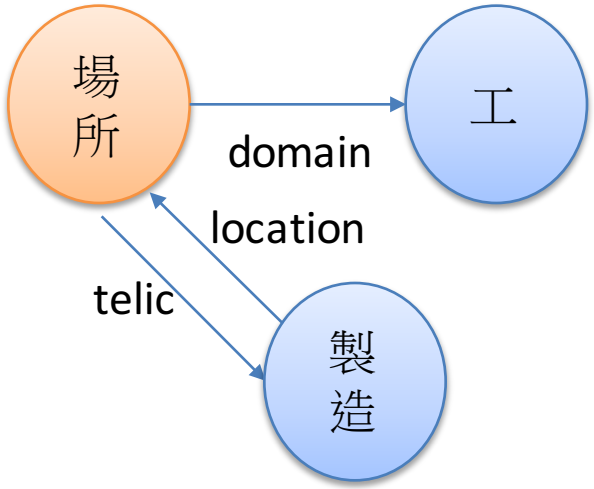
廣義知網(E-HowNet)

- 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},

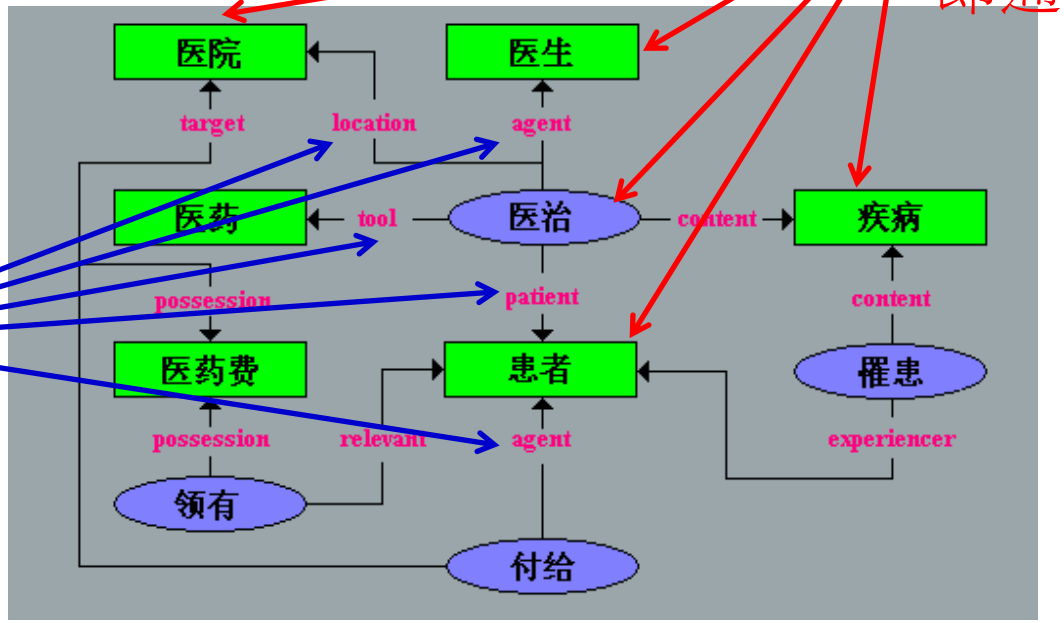
telic={produce|製造:location={~}}}

- 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—患者

- 國語辭典：病人
- 廣義知網：{humanl人:predication={or({illl病態:theme={~}},{doctorl醫治:patient={~}})}}



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={~}}}
- 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…
- 留學生 def:{human|人:predication={study|學習:agent={~},location={foreign|外國},domain={education|教育}}}
- 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}}

1. 相同的語義會有相同的表達式
2. 近似的語義會有近似的表達式
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

多層次表達

西瓜

- 基本概念式　　def1: {西瓜|watermelon}
- 展開為義原表達式　def2:{fruit|水果:

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}}}

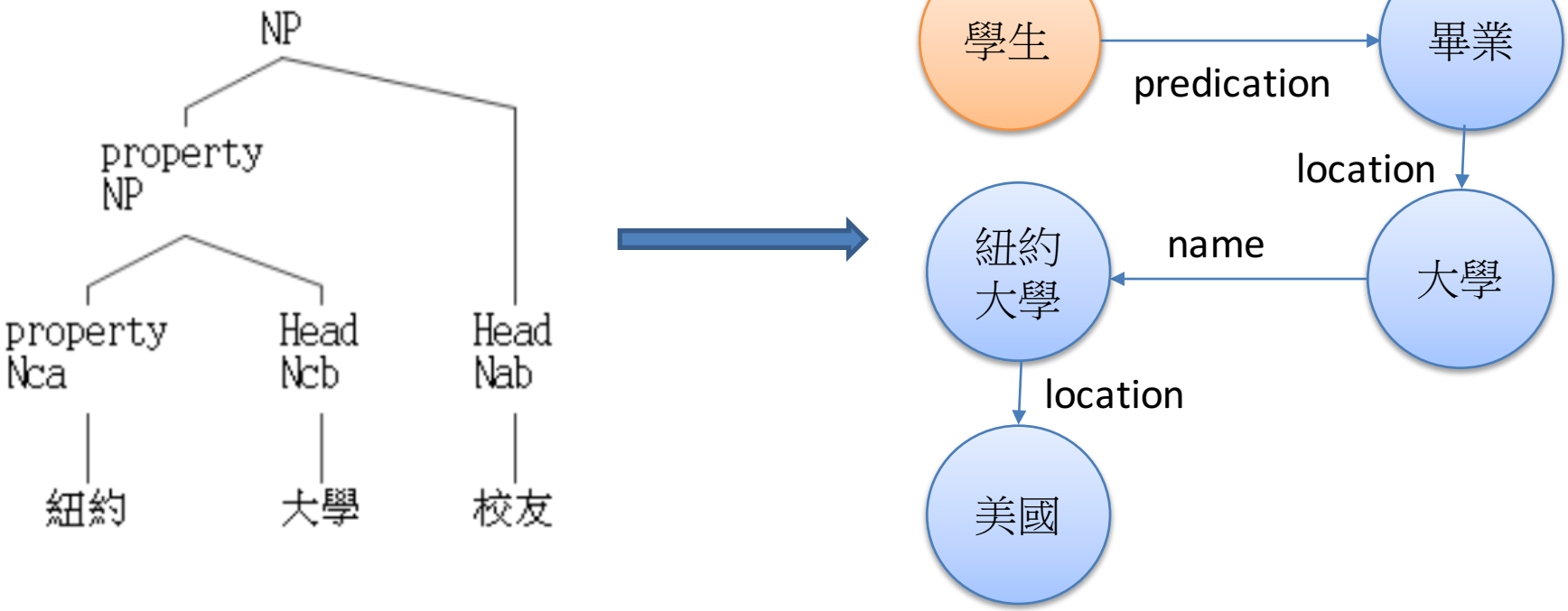
1. 義原：{英文|中文}
2. 基本概念：{中文|英文}
   基本概念保留更多的訊息
   讓表達式簡化、清晰
3. 多層次的語義表達式

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



Chinese NLU by CKIP

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

謝謝聆聽