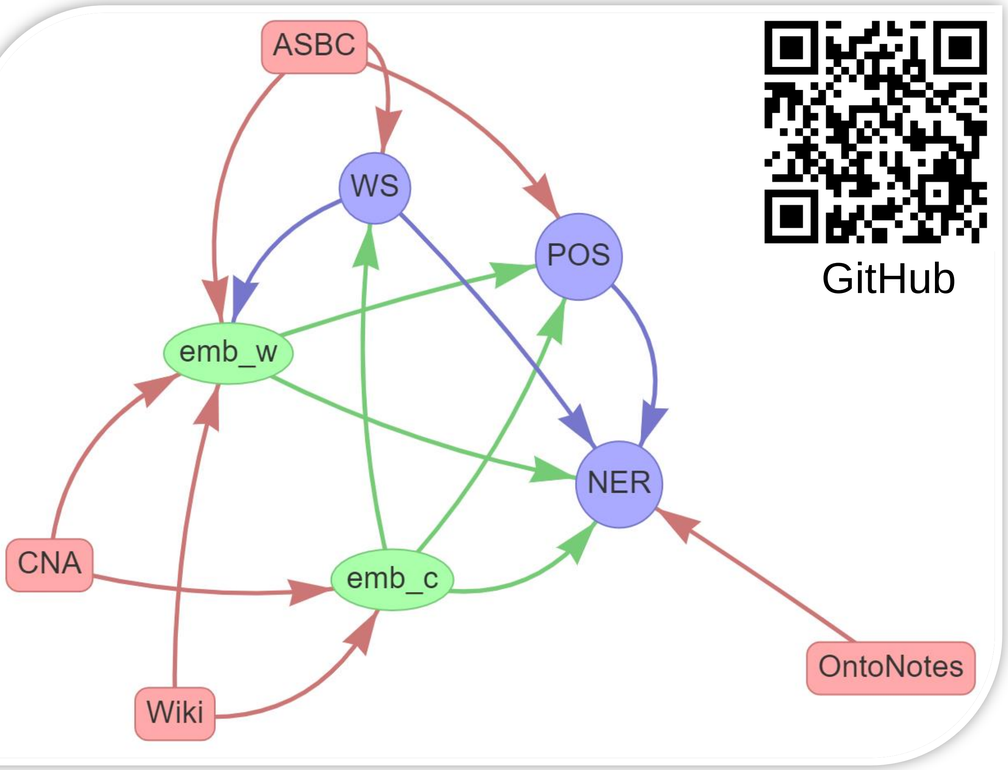
建構概念為本且具語義結合性的中文知識庫

馬偉雲助研究員團隊 中研院資訊所

CkipTagger: 開源的斷詞(WS)、詞性標注(POS)、實體辨識(NER)工具



GitHub

18類實體辨識

(人,事件,日期,地理,產品,組織,法律,時間,行政區,設施,
社會團體,語言,藝術品,錢,數字,序數,數量,百分比)

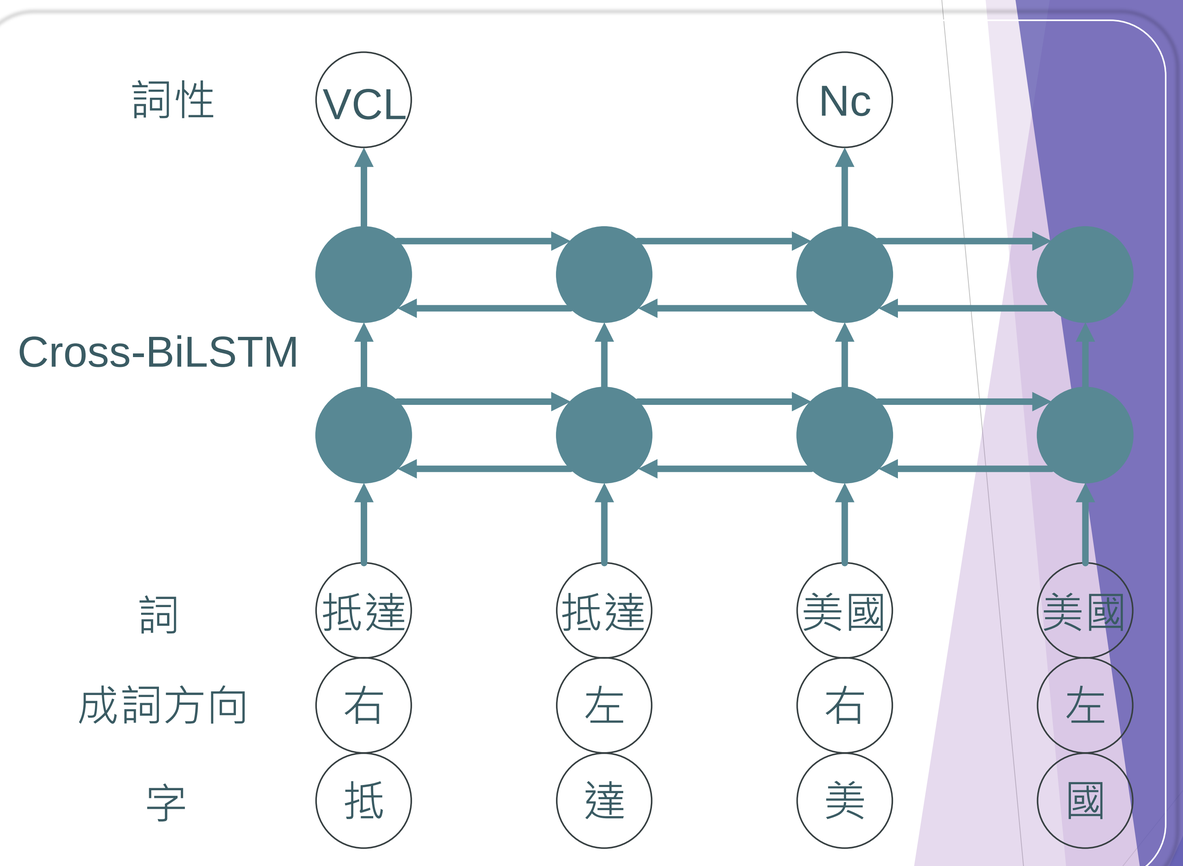
傅達仁PERSON今將執行安樂死，卻突然爆出自己20年前DATE遭緯來體育台ORG封殺，他不懂自己哪裡得罪到電視台。

61類詞性標注

(各類名詞,各類動詞,各類副詞,各類標點,...)

傅達仁(Nb) 今(Nd) 將(D) 執行(VC) 安樂死(Na) ，(COMMACATEGORY) 卻(D) 突然(D) 爆出(VJ) 自己(Nh) 20(Neu) 年(Nf) 前(Ng) 遭(P) 緯來(Nb) 體育台(Na) 封殺(VC) ，(COMMACATEGORY) 他(Nh) 不(D) 懂(VK) 自己(Nh) 哪裡(Ncd) 得罪到(VJ) 電視台(Nc) 。(PERIODCATEGORY)

詞性標注模型



語料

Corpus		Size	Training
CNA	中央社新聞	11億字	字向量/詞向量
Wiki	中文維基	4.6億字	字向量/詞向量
ASBC	中研院平衡語料庫	1633萬字	斷詞/詞性標注
OntoNotes	多重標記語料庫	151萬字	實體辨識

實驗分析

	ASBC-test				OntoNotes-Test
	WS			POS	NER
	Prec.	Rec.	F1	Acc.	F1
CkipTagger POS	--	--	--	***97.20***	--
CkipTagger WS+POS+NER	97.49	97.17	***97.33***	94.59	***77.98***
CKIPWS	95.85	95.96	95.91	90.62	--
Jieba_zhTW	90.51	89.10	89.80	--	--

$$*POS\ acc = \frac{\#\ predicted\ words\ with\ correct\ boundary\ and\ POS\ tags}{\#\ gold\ standard\ words}$$

論文

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. 2019. Remedying BiLSTM-CNN Deficiency in Modeling Cross-Context for NER. arXiv preprint arXiv:1908.11046.

AIntu 科技部臺灣大學人工智慧研究中心
科技部人工智慧技術暨全幅健康照護聯合研究中心
MOST JOINT RESEARCH CENTER FOR AI TECHNOLOGY AND ALL VISTA HEALTHCARE
MOST 科技部 Ministry of Science and Technology