結合斷詞、詞性標記、實體辨識的一站式中文處理開源套件 - CkipTagger

馬偉雲 助研究員
中研院詞庫小組主持人
2019/11/21

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界提供工具、資料、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。

CkipTagger Developers

李朋軒
(main developer)

傅子睿

馬偉雲

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

CkipTagger – An Open Source Toolkit

- GitHub:
  - https://github.com/ckiplab/ckiptagger
- Online Demo:
  - https://ckip.iis.sinica.edu.tw/service/corenlp
  - https://ckip.iis.sinica.edu.tw/service/ckiptagger

CkipTagger 特色

- 斷詞與詞性標記的表現進一步提升，並大幅超越結巴系統。

ASBC 4.0 測試集 (50,000句)

Tool	(WS) prec	(WS) rec	(WS) f1	(POS) acc
CkipTagger (詞庫小組新版系統)	**97.49%**	**97.17%**	**97.33%**	**94.59%**
CKIPWS (詞庫小組舊版系統)	95.85%	95.96%	95.91%	90.62%
Jieba-zh_TW (結巴系統)	90.51%	89.10%	89.80%	--

CkipTagger 特色

- 結合實體辨識（Named Entity Recognition，NER）
  - 目標是在文字資料當中，能夠辨識出感興趣的專有名詞(包含原本資料庫不存在的新專有名詞)，並自動標記正確的分類。
  - 目前CkipTagger能辨識11 類一般領域專有名詞及 7 類數量詞，包含：
    - 人名、團體、設施、組織、地理、地點、商品、事件、藝術品、法律、語言、日期、時間、比例、錢、數量、序數、數詞。

CkipTagger 特色

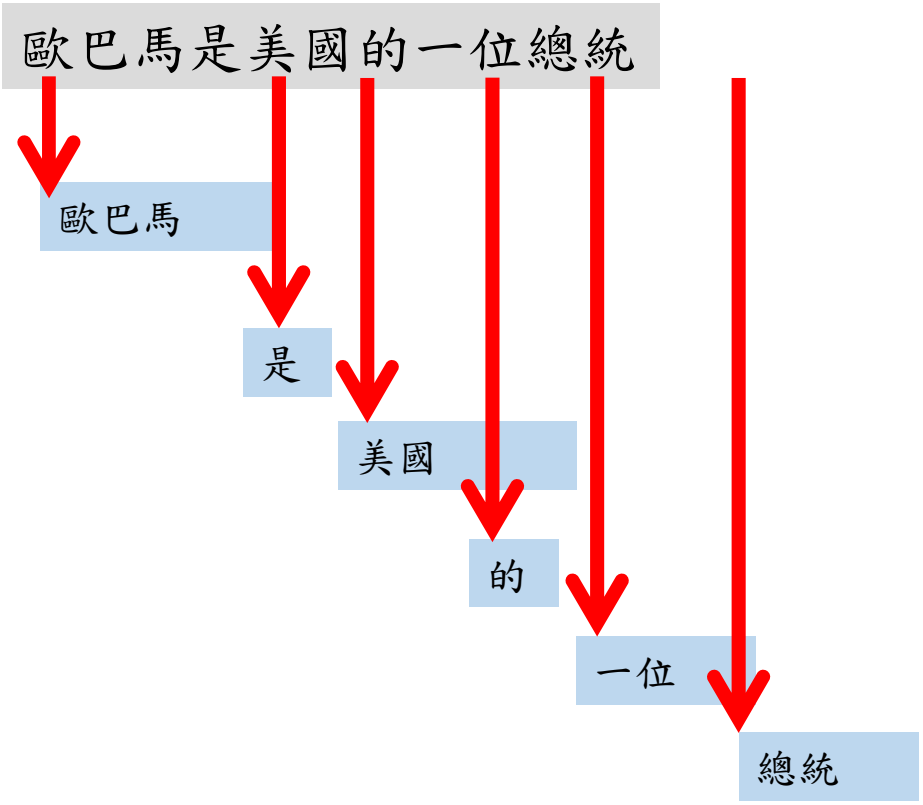
- 支援使用者自訂 參考/強制 詞典。
  - 從實際應用的角度，能夠支援使用者自訂詞典是一個相當重要的功能。一般而言，以字為標記單元的機器學習/深度學習的斷詞模型通常因為算法本身的特性，而難以提供使用者自訂詞典的功能。CkipTagger則克服了這個限制，雖是使用以字為標記單元的模型，但仍然支援使用者自訂詞典，包含參考詞典與強制詞典，且每個詞彙均可指定權重，讓使用者能根據自身的任務需求與領域，自行進行系統的客製化。
- 支援不限長度的句子。
- 不會自動 增/刪/改 輸入的文字。

斷詞技術分類

- Word-level approach
  - Maximum length (長詞優先)
  - JiaBa (動態規劃查找最大概率路徑)
- Character-level approach
  - Stanford (CRF)
  - CkipTagger (Cross-BiLSTM)
- Hybrid-level approach
  - CKIP Word Segmentation (CKIPWS)

Maximum length (長詞優先)

- One technique is Maximum length Word First

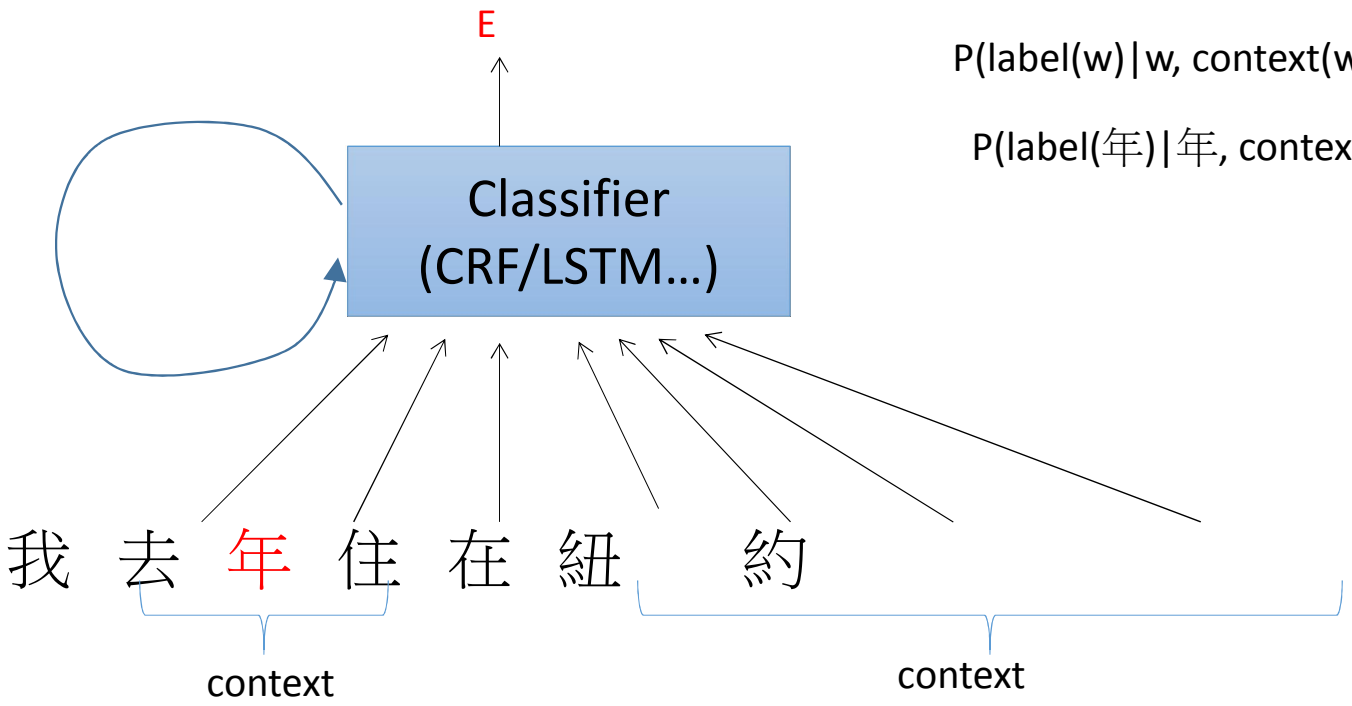


Character-level approach

- Character Sequence Labelling

我 去 年 住 在 紐 約
O B E O O B E

我 去年 住 在 紐約

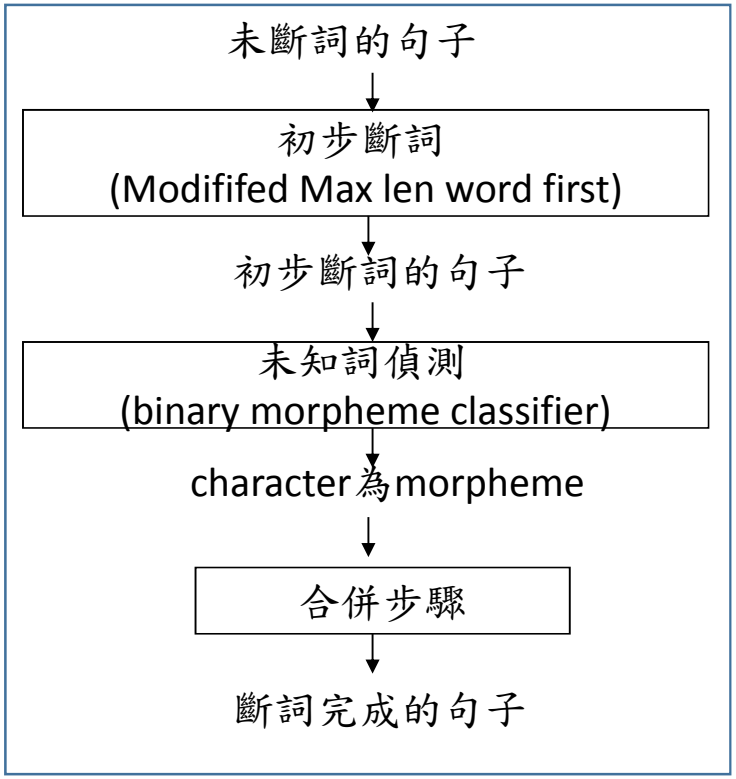


P(label(w)|w, context(w), previous words' labels)

P(label(年)|年, context(年), label(去), label(我))

CKIPWS (詞庫小組舊版斷詞系統)

- 流程



若能在營益率上維持良好成績

若 能 在 營 益 率 上 維持 良好 成績

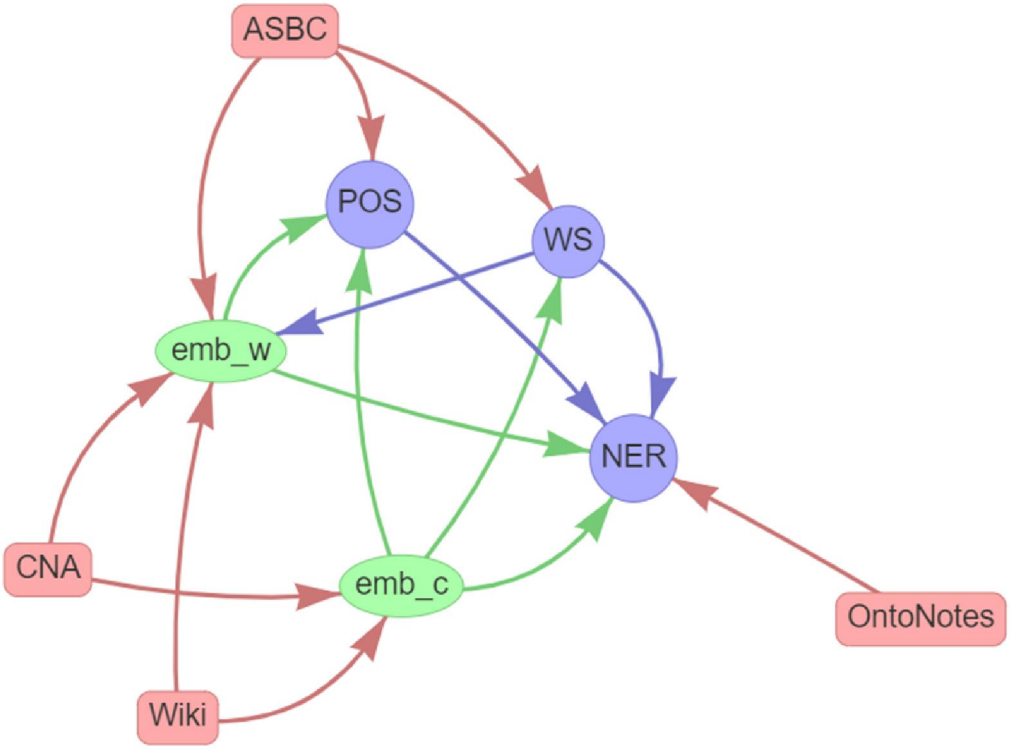
若 能 在 營(?) 益(?) 率(?) 上 維持 良好 成績

若 能 在 營益率 上 維持 良好 成績

Ma and Chen , "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction", SIGHAN of ACL, 2003
Ma and Chen , "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff", SIGHAN of ACL, 2003

Roadmap of CkipTagger (詞庫小組新版斷詞系統)

- Corpora
- Embedding
- Tools



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

Corpora

- Normalize
  - Unicode-normalization
  - Transform to ZhTW

- Datasets
  - CNA: Chinese Gigaword 5, CNA part
  - Wiki: Chinese wiki, 2019-05-20 pages-articles dump
  - ASBC: ASBC 4.0
  - OntoNotes: OntoNotes 5.0, Chinese part

	Sentences	Words	Characters	word/sent	char/sent	Sentence type
CNA	13,366,581	632,289,913	1,098,546,752	47.3	82.2	Paragraph
Wiki	5,557,141	247,714,633	461,862,002	44.6	83.1	Paragraph
ASBC	1,297,793	10,409,751	16,331,383	8.0	12.6	Clause
OntoNotes	46,905	958,345	1,515,151	20.4	32.3	Sentence

Embedding

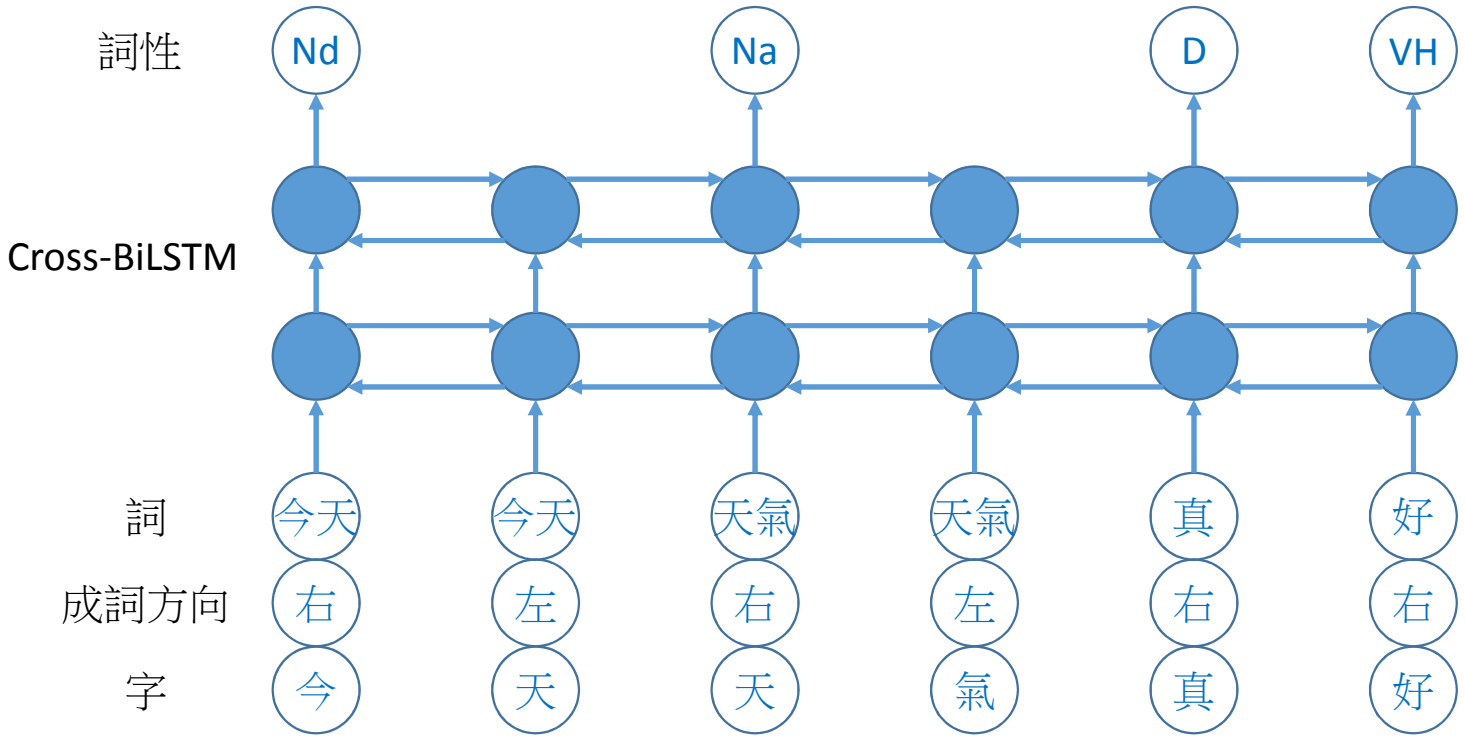
- Character
  - CNA + Wiki
  - 1,560,408,754 characters
  - [13136, 300]

- Word
  - CNA + Wiki + ASBC-train, segmented by WS
  - 890,414,297 words
  - [1355791, 300]
    - Excluding non-most frequent 30K and length>20

Tools

- WS
  - Cross-BiLSTM
  - Character-based
  - BI tagging
- POS
  - Cross-BiLSTM
  - Character-based with WS info
  - Only tag B characters
- NER
  - Cross-BiLSTM
  - Character-based with WS, POS info
  - BIOES tagging

POS Model



What is Cross-BiLSTM?

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

Baseline-BiLSTM-CNN

- Prevalently used as the core of NER sequence taggers
  - 2-layer BiLSTM
    - Extracting high-level features for each direction

$$\overrightarrow{H} = \overrightarrow{LSTM}_2(\overrightarrow{LSTM}_1(X))$$

$$\overleftarrow{H} = \overleftarrow{LSTM}_4(\overleftarrow{LSTM}_3(X))$$

$$H = \overrightarrow{H} \,||\, \overleftarrow{H},$$

  - Affine-Softmax tagging

$$s_t = H_t W_p + b$$

$$p_{ti} = \frac{e^{s_{ti}}}{\sum_{j=1}^{d_p} e^{s_{tj}}},$$

Baseline-BiLSTM-CNN

- XOR Limitation
  - Key and Peele (*work-of-art*)
  - You and I (*work-of-art*)
  - Key and I
  - You and Peele
  - Key -> *Inside* <- Peele
  - You -> *Inside* <- I
  - Key -> *Outside* <- I
  - You -> *Outside* <- Peele
  - Key+You -> *Inside* <- I+Peele
  - Key+You -> *Outside* <- I+Peele
- The 4 cases cannot be all tagged correctly (Contradiction)

Cross-BiLSTM-CNN

- Extracting high-level features for cross-context

$$\overrightarrow{H}^1 = \overrightarrow{LSTM}_1(X)$$
$$\overleftarrow{H}^3 = \overleftarrow{LSTM}_3(X)$$
$$\overrightarrow{H}^2 = \overrightarrow{LSTM}_2(\overrightarrow{H}^1 || \overleftarrow{H}^3)$$
$$\overleftarrow{H}^4 = \overleftarrow{LSTM}_4(\overrightarrow{H}^1 || \overleftarrow{H}^3)$$
$$H = \overrightarrow{H}^2 || \overleftarrow{H}^4$$

- Solving XOR limitation by additive non-linearity

Experiments on English NER

- Overall results

	OntoNotes 5.0			WNUT 2017		
	Prec.	Rec.	F1	Prec.	Rec.	F1
BiLSTM-CNN	86.04	86.53	86.28±0.26	-	-	-
CRF-IDCNN	-	-	86.84±0.19	-	-	-
CRF-BiLSTM(-BiLSTM*)	-	-	86.99±0.22	-	-	38.24
Baseline-BiLSTM-CNN	88.37	87.14	87.75±0.14	53.24	32.93	40.68±1.78
Cross-BiLSTM-CNN	88.37	**88.17**	88.27±0.17	**58.28**	33.92	**42.85**±0.99
Att-BiLSTM-CNN	**88.71**	88.11	**88.40**±0.18	55.82	**34.08**	42.26±0.82

WS, POS Performance on Chinese Benchmark

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

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

WS Performance with Dictionary

- Performance with target domain-specific dictionary
  - Preprocess: -0.2%
  - Hard post-process: -0.3%
  - Soft post-process: +0.2%

	ASBC-validate		
	Prec.	Rec.	F1
No dictionary	97.52	97.12	97.32
Preprocess	97.52	96.68	97.10
Hard post-process	97.17	96.90	97.03
soft post-process	97.80	97.21	97.50

- The final tool allows
  - A combination of hard + soft
  - Setting weights for each dictionary word

WS Speed

- Speed
  - (GPU) GTX 1080 Ti + (CPU) Xeon E5-2620 v4, using 3 cores

	Sentences	Words	Characters	word/sent	char/sent	sent/sec	word/sec	char/sec	Sentence Type
CNA	13,366,581	632,289,913	1,098,546,752	47.3	82.2	477	22,543	39,167	Paragraph
Wiki	5,557,141	247,714,633	461,862,002	44.6	83.1	410	18,271	34,066	Paragraph
ASBC-train	1,297,793	10,409,751	16,331,383	8.0	12.6	5,150	41,309	64,807	Clause

NER Performance

	Features	Embedding	OntoNotes Validate	OntoNotes Test
Old NER	CKIPWS	Morris	75.22	75.79
CkipTagger NER	CkipTagger WS+POS	CkipTagger	78.49	77.98

CkipTagger在自然語言處理的應用

- 聊天機器人
  - 詞庫小妍 (https://ckip.iis.sinica.edu.tw/project/chatq/)
- 輿情分析
  - http://learn.iis.sinica.edu.tw:9187/
- Much more
  - https://ckip.iis.sinica.edu.tw/demo/

詞庫小妍 – 陪你天南地北聊天的LINE機器人

- 詞庫小妍是中研院詞庫小組打造的一款不限領域、 能夠陪你天南地北聊天的聊天機器人, 於2019/10正式在LINE平台上對外開放, 只要掃描QR code, 就可以在您自己的手機上跟小妍聊天囉！
- 有四大聊天模式可供切換
  - 新聞嘴砲
  - 美妝達人
  - 文青女孩
  - 綜合模式

新聞嘴砲

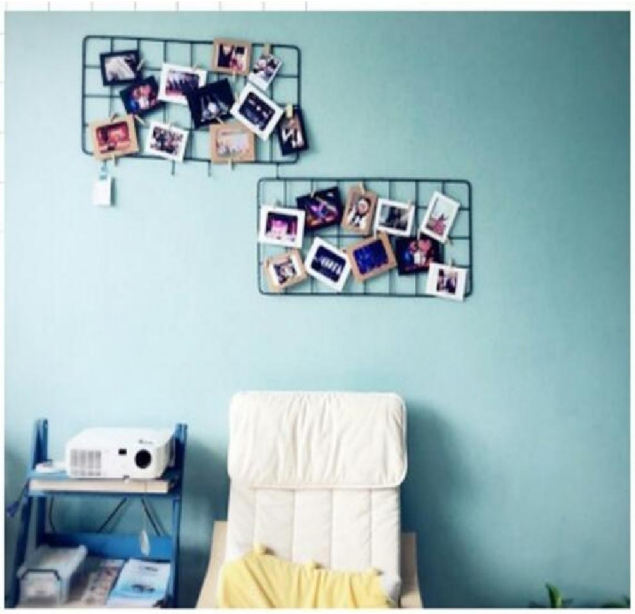

已讀
下午4:06
蔡英文
訪台南談反送中 蔡英文：台灣的命運由台灣人自己決定
自由時報 - 2019年08月31日
這則新聞
相關新聞
反送中》王丹：中共不願出兵 怕拉抬蔡英文選情
自由時報 - 2019年08月31日
這則新聞
相關新聞
還在質疑蔡英文學位？陳致中開砲：講到臭酸令人生厭煩！
自由時報 - 2019年08月31日
這則新聞
相關新聞

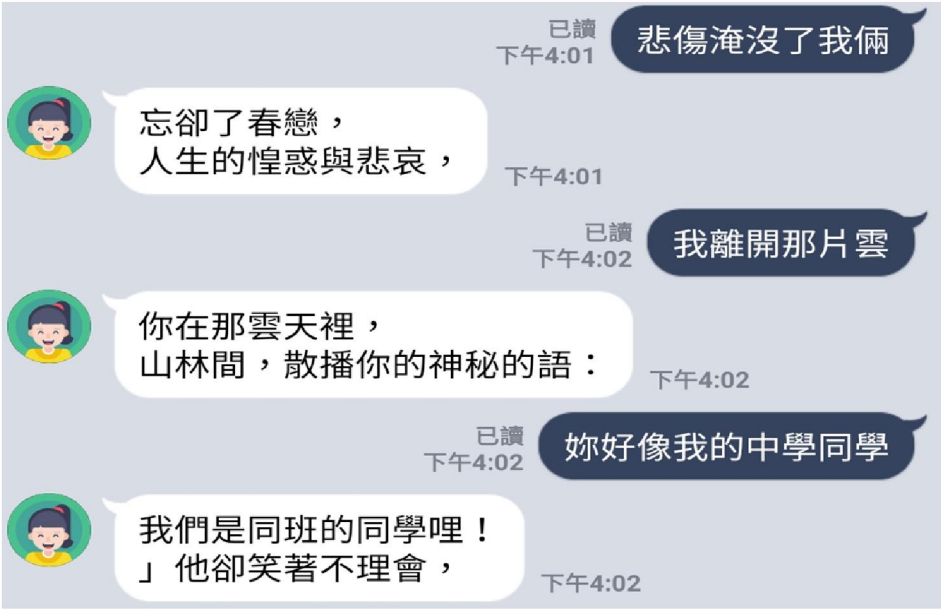

已讀
下午4:11
中美貿易戰
目前正逢中美貿易戰，中國產品無法銷往歐美
下午4:11
已讀
下午4:12
你打球
我打球是因為娛樂
下午4:12
已讀
下午4:13
你說醉話
我說的是事實！！！
下午4:14
已讀
下午4:14
我好累
你應該會沒精神吧！
下午4:14

美妝達人


嗨～ 跟我聊聊妳的化妝品，我會告訴妳很多道聽途說的資訊唷！^.<
下午11:06
你想知道哪個品牌的化妝品情報呢？
下午11:06
已讀
下午11:06
divinia 蒂芬妮亞
你想知道 divinia 蒂芬妮亞 的哪個產品的情報呢？
下午11:06
已讀
下午11:06
3D水感柔焦CC霜 SPF50
divinia
下午11:06
你想知道關於 3D水感柔焦CC霜 SPF50 的哪些事情呢？
下午11:06
控油
潤色
防曬
reset
選單

文青女孩




已讀
下午4:01
悲傷淹沒了我倆
忘卻了春戀，
人生的惶惑與悲哀，
下午4:01
已讀
下午4:02
我離開那片雲
你在那雲天裡，
山林間，散播你的神秘的語：
下午4:02
已讀
下午4:02
妳好像我的中學同學
我們是同班的同學哩！
」他卻笑著不理會，
下午4:02

Thank You!