應用 AutoNER 於社群網路中文歌手名稱辨識之研究

邱威誠 Wei-Cheng Chiu
資訊工程所, 國立中央大學
david22294@cc.ncu.edu.tw

張嘉惠 Chia-Hui Chang
資訊工程所, 國立中央大學
chia@csie.ncu.edu.tw

***Abstract*—本論文的研究主題為社群媒體上的中文命名實體辨識 (Named Entitiy Recognition, NER)。由於序列標記模型需要已標記的文本做為訓練資料，利用已知的字典透過自動標記的方式去產生訓練文本，是一種減少人工標記成本的方法。但在自動標記的過程中，可能會產生錯誤的標記而影響效能，如何解決這項問題是這項研究所面對的一大挑戰。我們參考 AutoNER 模型所使用的 Tie 和 Break 標記替代傳統的 BIEOS 標記方式，做為本論文模型的基礎；同時考量中文缺少英文之字詞分隔，限制語意的理解，因此我們額外地加入了中文斷詞的資訊，提高標記中 Tie 的比例，借以輔助序列標記模型的訓練。實驗結果顯示，在新的資料標記中，相較於完全比對 (exact match) ，加入約略比對 (approximate match) 對模型 F1 的效能提升 18% ；比起採用 BIEOS 標記之 CRF 架構下的模型，AutoNER 架構下的模型在 F1 的效能提升 9%。**

I. Introduction

對於唱片業、娛樂業而言，挖掘有潛力的藝人、了解其市場的關注度，對歌人的培植、投資策略有其重要性。而機器學習與文本分析的進展，使得企業可以透過社群網路之意見探勘得知市場反應度，不僅可以幫助企業了解民眾對於藝人的反應，做為產業行銷與決策的考量。

本論文目標為社群媒體歌手及團體名稱之命名實體辨識不同於較為嚴謹的報章雜誌，社群媒體流行不少因為時事而創造出來的詞彙。例如「飯」這個詞在社群媒體上代表的是英文字 Fan ，所以常會看到大家會用「XX 飯」去稱呼自己為某個藝人的粉絲。這類社群媒體上特有的用語都是造成機器在模型學習上困難的原因；除此之外，對於同一位歌手可能會用多種暱稱來稱呼，例如：「周董」指的是台灣歌手周杰倫、「Angela」指的是台灣歌手張韶涵等。除此之外，團體名稱又更加多樣化，例如：「四個朋友」為擎天娛樂所栽培的學生歌手樂團；「有感覺」為台灣獨立樂團。「四個朋友」與「有感覺」同為團體名稱，但是也是十分口語化的表達。因此基於字典對訓練資料進行標記時，常會有誤標非歌手的偽陽性標記 (False Positive) 或是漏標歌手實體名詞的狀況 (偽陰性標記)，造成訓練文本本身就有較高的錯誤率，這也是利用已知實體詞典做自動標記時必須要面對的問題。

為了降低在自動標記中錯誤標記造成的影響，我們參考了 Shang 等人 [11] 所提出的 AutoNER 模型及特殊的 Tie 和 Break 資料標記方式。不同於常見的 BIO 或是 BIEOS 序列標記方式，AutoNER 模型在架構上是利用 Multi-task learning 去做模型的設計，將 NER 任務方式拆成兩個輸出去做預測，分別是片語預測和類別的預測，在這邊比較特別的是片語預測是利用 Tie 和 Break 兩個標籤去決定兩個字之間的關係，此標記能夠有效降低錯誤標記的噪音所造成的影響。

由於 AutoNER 模型本身是針對英文任務去做設計，為了能夠適應中文任務，我們也對 AutoNER 模型去做調整。首先，中文不同英文在沒有明確的斷詞方式，同樣一句話可能因為斷詞方式不同而產生不同意思，因此在片語預測的部分，不同於 AutoNER 單純針對標記對像做片語預測，我們將此任務層的任務從單純的片語預測改為中文的斷詞預測，其次，我們將 AutoNER 模型中的 word embedding layer 改成採用 pre-train BERT 模型再進行 Fine-Tuning。AutoNER 的另一個特點在於，採用一個特殊的未知類別標記，針對約略比對 (approximate match) 的字詞給予未知類別標籤。例如：「少時的歌曲很棒」和「少時離家老大回」兩句話中都含有「少時」兩個單詞，然而前者指的是「少女時代」的韓國偶像團體，後者的「少時」則是指年輕的時候，基於上述例子我們給予部分比對的字詞未知類別標籤，並在計算損失函數時忽略其標籤的計算，避免模糊字詞對於模型的影響。

在實驗的部分，我們針對 1. 自動標記的方式、2. 將片語預測改為中文斷詞預測、3. AutoNER 架構與 CRF 架構比較，三種主題去做效能上的比較。結果顯示，使用未知類別標記處理部分比對字詞、以及將片語預測改為中文斷詞預測，對於任務的整體 F1 效能上可以提升 18% 及 3% 效能。而在本論文針對中文任務所修正的 AutoNER 在整體效能上也有著較好的效能。

II. Related Work

A. *Multi-task Learning*

多任務學習有很多形式，像是聯合學習 (Joint Learning) ，自主學習 (Learning to Learn) 等，針對多個目標函數做最佳化的架構都能稱為是多任務學習。多任務學習通過使用蘊含在相關任務的監督信號中的領域知識來改善泛化效能 (generalization ability) 。在機器學習中，多任務學習被視為一種 inductive transfer，透過 inductive bias 的引入來改善模型。舉例來說，L1 正規化是一種常見的歸約偏置。在多任務學習中，歸約偏置是由多種任務來提供，結果上這能減少模型過擬合的風險。

多任務學習在深度學習中常用的方法有兩種，隱藏層參數的硬共享 (Hard Parameter Sharing) 與軟共享 (Soft Parameter Sharing) 。在軟共享機制中，每個任務都由自己的模型，自己的參數；而硬共享機制 [2] 是多任務學習中最常見的一種方式，它可以被應用到任何任務及任何隱藏層上，且保有任務相關的輸出層。根據 Baxter 在 Multitask sampling 的貝氏及資訊理論模型 [1]，這些共享

MOST108-2622-E-008-012-CC2

A. Training Data without CWS

Token	吳	亦	凡	，	倒	數	亦	凡	與	你	的	時	間	距	離	，	開	啟	吳	小	爺	大	好	時	的	人	生
Type	A	A	A	N	N	N	U	U	N	N	N	N	N	N	N	N	N	N	U	U	U	N	N	N	N	N	N
Chunk		T	T	B	B	B	U	U	B	B	B	B	B	B	B	B	B	B	U	U	U	B	B	B	B	B	B

B. Training Data with CWS

Token	吳	亦	凡	，	倒	數	亦	凡	與	你	的	時	間	距	離	，	開	啟	吳	小	爺	大	好	時	的	人	生
Type	A	A	A	N	N	N	U	U	N	N	N	N	N	N	N	N	N	N	U	U	U	N	N	N	N	N	N
Chunk		T	T	B	B	T	B	T	B	B	B	B	T	B	T	B	B	T	B	U	T	B	T	B	T	B	T

圖 1. 資料自動標記範例：A 原始方法（無斷詞），B 新方法（考慮斷詞）

參數共擬合風險（介於 0 1 間之數值）的階數為 N，而 N 為共同任務的數量，這使多任務學習相較單任務學習過擬合的風險還要小，越多任務同時學習，我們的模型就能捕捉到越多任務的共同表示法，從而降低我們在原始任務上的過擬合風險。

B. *中文命名實體辨識*

命名實體辨識 (NER) 的目的在於從非結構化中的文本擷取實體並根據其類別進行標記，常見的命名實體類別有人名、地名、組織等特徵字詞。現今 NER 模型的設計上，大多採取序列標記方法在輸出層上使用條件隨機場 (CRF) 作為輸出層，並利用 BIO 或是 BIOES 來標記命名實體的開始（B）、中間（I）、結束（E）、或其他（O）。有些模型同時考量多種命名實體之辨識，因此尚需預測類別，如 [3], [5]–[7] 等。在上述的論文中，多是使用監督式 (supervision) 的方式去做訓練，此種方式需要大量的人力去做訓練資料的標記。

為減少人工標記成本，近年來則有採取已知實體列表進行自動標記方式的提出 [4], [8], [10], [11], [13]，此類遠程監督式學習 (Distant Supervision) 希望在花費最少人力的情況下達到接近監督式學習的效果。[4] 除提出自動標記的概念，並採用 Tri-training 架構提升序列標記效能；[13] 提出提高自動標記資料的品質的方法，其使用少量人工標記資料混合自動標記資料，透過強化學習的方式，從自動標記資料中篩選出品質較高的標記資料去做模型的訓練；[10] 中提出 teacher-student 的訓練架構，運用 pre-train BERT 中蘊含的語意知識來輔助模型的訓練；[11] 中提出 Tie/Break 序列標記方法，搭配 Unknown 類別處理不確定性的標記，比較 MLP 及 CRF 兩種最終層架構之效能。本文延伸 [8] 對歌手辨識之研究，應用 AutoNER [11] 改善效能。

III. Methodology

A. *自動標記方式*

AutoNER 原始標記方法，會根據目標實體的字典將文本中完全比對的實體給予符合實體類別的標記，再來針對與部分比對的字詞，例如：「杰倫」之於字典中的「周杰倫」、「大雷」之於字典中的「何大雷」等，標記 Unknown 標籤，藉此去降低像是「大雷」這類可能是人名也有可能是指大自然現象的模糊字詞對訓練所造成的影響，標記結果如圖 1(A) 所示，chunk 為 Tie 或 Break 或 Unknown；type 為類別 Artist 或非藝人 N 或 Unknown。每一個 chunk 標記代表目前的字是否與「前者」相連。以圖中的「吳亦凡」為例，由於「吳」為開頭首字，所以沒有標記，而「亦」和「凡」能與前者組成一詞「吳亦凡」，所以兩者皆標記為 Tie 標籤，其他均標記為 Break。由於此標記方法只針對目標實體標記 Tie 標籤，因此 Tie/Break 比例會差別很多。這種的標記方式是以中文「字」為基礎去做自動標記，只要是部分比對的字詞不論是 type 或是 chunk 標記都會給予 unknown 標記。

第二種標記方式是以中文「詞」為基礎，這邊我們使用到了中研院的中文計算語言研究小組所提供的 CKIP [9]。先是利用字串比對，將完全比對的實體給予標記 Artist，再將標記完的文章丟入 CKIP 做中文斷詞，之後將斷詞後的結果中，未標記的單字詞再與實體字典做部分比對，將部分比對成功的實體標記上 unknown 標籤，其餘則標記成 None。最後將 CKIP 斷詞後的資訊加入模型 chunk 標記（圖 1(B)）。在此新標記方法中，我們不僅限於將目標實體標記為 Tie，對於經過 CKIP 分詞過後的多字詞都會給予 Tie 標記。當有兩個連續的詞類別 Type 皆為 unknown 時，會給予兩詞之間的 chunk 標記上 unknown 標籤。

B. *模型架構*

我們使用 pre-trained Bert 作為嵌入層 (embedding layer) 且同時做為 Multi-task 架構中的共享層 (shared layer)，在訓練時會同時做 fine-tuning 的動作；在個別的任務層 (task specific layer)，除了 AutoNER 論文中所使用的 Highway layer 之外，由於我們認為中文在片語預測（斷詞預測）上難度較英文高，需要較複雜的模型去做應對，所以我們嘗試在下游層加入 BiLSTM 層，觀察對於較複雜的資料集，增加模型複雜度是否能提高模型總體的效能（圖 2）。

輸入層 (Input Layer)

將經過 Bert 本身提供的 tokenizer 處理後的句子作為輸入，tokenizer 針對中文任務是以字元為單位 (character based) 將字轉換為 bert 字典中對應的編號，表示符號為 $S^p = \{c_1, c_2, ..., c_n\}$，$S^p$ 代表訓練與料庫中的第 p 個句子，c_i 表示句子的第 i 個字元的編號，總共有 n 個字元，而每個句子會有兩組針對每個字元的標記 $Y_p^{chunk} = \{C_1, C_2, ..., C_n\}$ 和 $Y_p^{type} = \{T_1, T_2, ..., T_n\}$ 分別為 chunk 和 type 的標記（如圖 1）。

嵌入層 (Embedding Layer) 與共享層 (Shared Layer)

本實驗的模型使用 Bert 同時作為模型的嵌入層以及 Multi-task 下的共享層。將輸入層 S^i 中的詞彙透過 Bert 模型轉換成數值型向量，轉換函數為 $x_i = BERT[c_i]$，轉

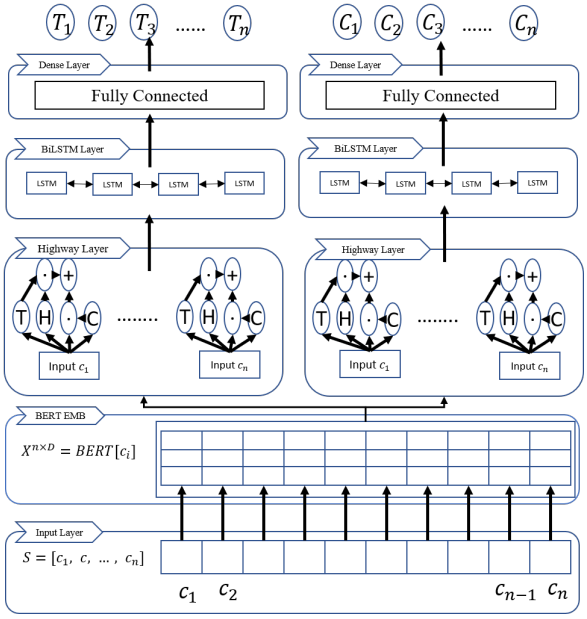

圖 2. Bert-Highways-BiLSTM 模型架構

換後的數值 $x_i \in R^D$ 是 c_i 轉換後的字元向量，D 表示向量維度 (Bert 預設為 768)，若 tokenizer 的字典中沒有該字，則以 UNK 所代表之向量取代。

個別任務層 (Task Specific Layer)

在這個模型中，我們使用到了兩個模型架構於個別任務層，分別為堆疊的高速網路結構 (Highway) [12]，加上雙向短期記憶循環神經網路 (BiLSTM)，根據 [12] 所述，高速網路結構運用類似 LSTM 的門控機制，利用兩個門，變換門 (Transform gate) 和進位門 (Carry gate) 來控制當層的輸入資訊，Transform gate 控制上一層的資訊有多少要經過非線性轉換後才能進入下一層，Carry gate 控制上一層資訊有多少能不經過非線性轉換直接進入下一層。透過此結構，我們能解決網路堆疊和 BERT 本身 768 維度的字向量造成參數量過多，導致梯度資訊回流受阻，造成網路訓練困難的問題。

在此層，首先我們將經過嵌入層所得到的字向量 x_i 作為 Highway 層的輸入，得到輸出 y_i，轉換函數公式為 $y_i = Highway(x_i)$ (公式 1)，這邊 H 為一非線性轉換運算、T 為 transform gate 運算、C 為 Carry gate 運算，W 則為可訓練之權重，為了計算方便，我們假定 $C = 1 - T$，公式 1簡化成公式 2。之後再將結果輸入到 BiLSTM 層，得到 $\overrightarrow{h_i}$ 和 $\overleftarrow{h_i}$ 並將兩者做相加得到 BiLSTM 最後的結果 $h_i = \overrightarrow{h_i} + \overleftarrow{h_i}$。

$$y_i = H((x_i, W_H) \cdot T(x_i, W_T) + x_i \cdot C(x_i, W_C) \quad (1)$$

$$y_i = H(x_i, W_H) \cdot T(x_i, W_T) + x_i \cdot (1 - T(x_i, W_C)) \quad (2)$$

輸出層 (Output Layer)

輸出層的部分，這邊我們將個別任務層的向量分別輸入進各自的全連接層後作 Softmax 後，將結果作為最後輸出，公式分別為 $\bar{T}_i = Softmax(W^T * h_i + b^T)$ 和 $\bar{C}_i = Softmax(W^C * h_i + b^T)$，$\bar{T}_i$ 和 $\bar{C}_i$ 的維度分別為 NER 的類別以及 2(Tie 和 Break)。

損失函數 (Loss Function)

這邊針對兩個任務皆採用 CrossEntropy 作為訓練時的損失函數，如公式 3和公式 4。在 multi-task learning 的結構下，兩者的權重為 1 : 1 ，因此在訓練時模型的最終 loss 為兩者相加 (公式 5)。

$$Loss^T = \sum_{1}^{n} CrossEntropy(T_i, \bar{T}_i) \quad (3)$$

$$Loss^C = \sum_{1}^{n} CrossEntropy(C_i, \bar{C}_i) \quad (4)$$

$$Loss = Loss^T + Loss^C \quad (5)$$

IV. Experiment

我們準備了含有 7,700 筆歌手藝人名稱的字典做為查詢詞，從 PTT 社群網路檢索含有目標實體的句子，共 61,685 句進行自動標記，根據不同標記方式會有不同筆數的 Entities 及 mentions 數據 (表 I)，部份比對多出完全比對之量即為標記為 Unknown 的 Entities 及 mentions 數。測試資料同樣是從社群網路媒體上蒐集下來的文章，並經過人工標記出答案 (如表 II)。

表 I
NER 訓練資料集 (自動標記)

Training Dataset		
# Seeds	7,700	
# Sentence	61,685	
Labeling Strategy	完全比對	部分比對
# Match Entities	4,987	10,323
# Mentions	81,562	108,823

表 II
NER 測試資料集 (人工標記)

Testing Dataset	
# Sentence	51,524
# Entity Mentions	26,067
# Distinct entities	4,327
OOV mentions (#/%)	17,695(68%)
OOV entities (#/%)	3,577(83%)

A. *評估方式*

我們使用 Precision/Recall/F1 來評估 NER 模型的效能。在評分方式上採「全對給分」的方式，歌手名稱必須完整預測出來才會被視為正確預測 (True Positive) 。由於在社群媒體中，粉絲們多使用暱稱來稱呼藝人歌手，但我們手中現有的藝人歌手字典多為正規名稱，導致測資中的 OOV 實體佔了極大的比例，也因此在此任務中，我們相較於 InV 的效能，我們更注重於 OOV 實體辨識的效能。

B. *自動標記策略*

首先我們比較完全比對以及部分比對兩種標記策略對於 AutoNER 模型效能的影響。從表 IV 中能看出，針對文本部分比對的實體應用 unknown 標記去忽略損失函數的計算，能明顯提升 NER 在整體、InV 和 OOV 的效能。

表 III
NER 模型效能評估表

		Overall			InV			OOV		
Data+Structure	Model	P	R	F1	P	R	F1	P	R	F1
Exact Match	Lattice	43.6%	47.7%	45.5%	68.9%	98.1%	81%	36.5%	25.7%	30.2%
	Bert-CRF									
Partial Match	BERT-CRF	50%	51%	51%	74%	93%	82%	40%	29%	34%
	BERT-GCNN-BiLSTM-CRF [8]	31%	72%	43%	57%	98%	71%	35%	59%	44%
AutoNER + WS	BERT-Highway	35%	66%	46%	57%	99%	72%	38%	51%	43%
	BERT-Highway-BiLSTM	41%	71%	52%	58%	98%	73%	44%	58%	50%

表 IV
NER 標記策略 F1 效能

Strategy	Overall	InV	OOV
完全標記	25%	52%	30%
部分比對	43%	70%	42%

C. *加入中文斷詞後的影響*

其次我們比較在 AutoNER 的 Tie 和 Break 的預測層中，使用只針對目標實體做預測的片語和中文斷詞預測，兩者之間效能的差異。如表 III所示，針對較為複雜的中文任務將片語預測任務改為中文斷詞任務能夠提升模型整體的準確度。

D. *AutoNER 架構與 CRF 架構之比較*

最後，根據上面實驗的結果，我們使用 Approximate Match 的訓練資料以及將片語預測改為中文斷詞預測做為 AutoNER 模型的基本設定。我們與傳統採用 BIEOS 標記的 BERT-CRF 模型以及 Bert-GCNN-BiLSTM-CRF [8] 模型做比較。同時我們考量 AutoNER 模型加入 BiLSTM 對於中文歌手辨識任務的影響。結果如表 V所示，雖然使用 BERT-CRF 有著 51% 的 F1，然而在 OOV 的部分 BERT-CRF 效果卻不盡理想，在 AutoNER 架構下的 BERT-Highway-BiLSTM 不但在 OOV 上有所提升，在整體 F1 上也小幅度的超越 BERT-CRF ，在整體的效能上有著最優異的表現。

表 V
加入中文斷詞後效能比較

Bert Highway		片語預測	斷詞預測
Overall	Precision	32%	35%(+3%)
	Recall	66%	66%
	F1	43%	46%(+3%)
InV	Precision	55%	57%(+2%)
	Recall	99%	99%
	F1	70%	72%(+2%)
OOV	Precision	35%	38%(+3%)
	Recall	51%	51%
	F1	42%	43%(+1%)

V. Discussion

綜合上述實驗，AutoNER 的架構確實能夠改善在做自動標記時容易遇到的困難，藉由使用 chunking 的 Tie/Break 標記以及 unknown 標籤取代 BIEOS 之標記，可以減少不必要的損失函數計算，不僅對於英文，也對於中文任務也都是有效的實作。同時我們將原本的片語預測任務改為中文斷詞任務，以及加入 BiLSTM 層去增加模型複雜度，能使其模型更加適應中文任務，從結果上看來這部分的假設確實能夠有效的提升模型的效能。

References

[1] Jonathan Baxter. A bayesian/information theoretic model of learning to learn via multiple task sampling. In *Machine Learning*, pages 7–39, 1997.

[2] Richard Caruana. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48. Morgan Kaufmann, 1993.

[3] Jason P.C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370, 2016.

[4] Chien-Lung Chou and Chia-Hui Chang. Named entity extraction via automatic labeling and tri-training: Comparison of selection methods. In *Information Retrieval Technology - 10th Asia Information Retrieval Societies Conference, AIRS*, volume 8870 of *Lecture Notes in Computer Science*, pages 244–255. Springer, 2014.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[7] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016.

[8] Gui-Ru Li and Chia-Hui Chang. Semantic role labeling for opinion target extraction from chinese social network. In Francesca Spezzano, Wei Chen, and Xiaokui Xiao, editors, *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining*, pages 1042–1047. ACM, 2019.

[9] Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. Why attention? analyze bilstm deficiency and its remedies in the case of NER. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 8236–8244. AAAI Press, 2020.

[10] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1054 – 1064, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[12] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *CoRR*, abs/1505.00387, 2015.

[13] Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. Distantly supervised NER with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.