The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 62-71

預訓練詞向量模型應用於客服對話系統意圖偵測之研究

Study on Pre-trained Word Vector Model Applied to Intent Detection of Customer Service Dialogue System

陳冠宇　Guan-Yu Chen
郭敏楓　Min-Feng Kuo
楊宗憲　Tsung-Hsien Yang
陳俊勳　Chun-Hsun Chen
廖宜斌　I-Bin Liao

中華電信研究院 巨量資料研究所
Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan, R.O.C.
robinchen@cht.com.tw
kmf0822@cht.com.tw
yasamyang@cht.com.tw
jeffzpo@cht.com.tw
snet@cht.com.tw

摘要

近年來對話商務的概念在各大科技巨頭間興起，人機互動方式由圖形化介面轉向對話交互介面的方式。因而自然語言成為人機互動介面的關鍵因子。然而教導機器要如何與人類溝通，以完成一項具體任務是相當有挑戰性的。其中一個需要克服的困難是自然語言理解，包含如何辨識使用者在詢問何種問題及如何取得文字間隱藏的資訊。讓機器了解使用者的問題意圖及資訊是相當重要的。本研究主要是針對去識別化後的中文客服對話資料，利用深度學習模型以達到辨識使用者意圖。為了更有效處理中文未知詞以及減少錯誤辨識，本研究比較不同預訓練詞向量模型與深度學習模型來辨識使用者意圖。相較於使用隨機詞嵌入，使用 BERT-WWM-Chinese (BWC)模型的正確率提升近 10%。這表示 BWC 模型產生的向量更能抓住用戶問句字詞間的語意關係。使得語意相近的字詞能產生近似的向量進而提升使用者意圖辨識的準確率。

Abstract

In recent years, the concept of dialogue business has arisen among major technology giants, and the way of human-computer interaction has changed from a graphical interface to a

dialogue interaction interface. Therefore, natural language has become a key factor in the human-computer interaction interface. However, teaching the machine to communicate with humans to accomplish a specific task can be quite challenging. One of the difficulties that needs to overcome is natural language understanding, including how to identify what questions users are asking and how to get information hidden between words. It is important to let the machine know the user's intentions and information.

The dataset of this study is collected from the dialogue of customer service materials. User's intents are recognized by deep learning models. In order to process Chinese unknown words more effectively and reduce false recognition, this study compares different pre-training vector models and deep learning models to understand user's intents. Compared with the use of random word embedding, the correct rate of using BERT-WWM-Chinese (BWC) model is improved by nearly 10%. It shows that the semantic vector generated by BWC model can better represent the relationship between user's words. The recognition rate of user's intent raises because similar vectors can be generated from similar words.

關鍵詞：對話系統，對話行為，深度學習，預訓練詞嵌入模型，注意力機制

Keywords: Dialogue System, Dialogue Act, Deep Learning, Pre-trained Word Embedding, Attention.

一、緒論

在科技不斷進步的今日，電腦、智慧型裝置、網路資訊服務在人類生活中日益扮演著重要的角色，人跟機器之間已有著密不可分的關係。也因為近年來人工智慧發展快速，更多應用深度學習技術來精進傳統機器學習方法的技術。

在許多情景下，對話用戶介面比圖型用戶介面更加自然及高效率，加上智慧型手機之發展，相較於撥打傳統語音客服，年輕人開始轉向使用文字對話介面與客服互動。傳統語音客服的各項任務，皆可經由文字對話描述來實現互動服務。許多公司及個人都嘗試著架構專屬的聊天機器人(Chabot)。然而，聊天機器人的功能不僅僅侷限於聊天，能夠以對話的方式來協助人類完成各式各樣目標才是我們真正想要的人工智慧。

在對話系統中，一般系統的輸入可以為語音或是文字，基於語音系統架構以 Steve Young 提出的架構最為典型[1]，而以文字為輸入的系統架構可以分為三大部分，首先是自然語言理解 (Nature Language Understanding, NLU) ，接著為對話管理 (Dialogue Manager,

DM) ，最後為自然語言產生 (Nature Language Generation, NLG) 。

機器人對話系統大致上可分為 2 種，分別是：

1. 聊天型對話系統：其不需要理解問題，只需要依據模型預測給予答案即可。

2. 任務型對話系統：需要分析問題意圖並給予預先建立之領域知識庫所定義的答案，其實作困難度相較於聊天型對話系統更高，因為回覆必須精準解答問題。

以實用性而言，任務型對話系統遠較聊天型對話系統要高，因為其可以給予用戶較有資訊意涵之回應。其衡量指標在於讓用戶自助服務率提高、並提升用戶滿意度，這包含對話輪次越少就能找到答案越好、可支援的意圖越多越好、回覆的內容越精準確實越好。在傳統文字客服系統上，用戶進線與客服交談以取得對應的解答。但是受限於公司成本考量、客服人力有限且每個值機員依其專業度，回覆內容之品質不一；若是能將大部分較單純、基本的問題，轉交給機器人來回答，那麼就可以節降公司不少值機成本，且可確保回覆內容一致，此外，用戶也可以節省掉等待值機員空閒、打字的時間，快速取得其需要的服務。該如何應用這些客服的文字記錄，研發出針對特定領域能自動問答的對話系統，是本研究的研究動機。

真實的客服對話系統中，使用者不管是表達句子方式或是詢問內容是否為該公司的業務，都不會受到任何限制，想問甚麼就問甚麼。輸入的句子較為口語化，因此經常有省略標點符號，打錯字，表達用詞上、文法上的錯誤，或是詢問非該領域的內容，導致無法使用預先定義的 ontology，在語意理解上也相當困難。因為任務導向型對話系統需要準確的理解語意，如何訓練機器去理解人類口語對話的語意，是相當重要的關鍵部分，如能有效解決此問題，將對自然語言理解能有更大進展。近年來，隨著深度學習技術發展迅速，國內外已經在對話系統有許多相關研究，目標都是在建立一套具有智慧的對話系統[2-5]。針對自然語言理解的研究，使用真實中文對話資料集的研究相當稀少，本文將比較不同深度學習模型是否可以在真實中文客服資料集中正確的辨別對話行為。

二、研究方法

一個說話者在對話中所要傳遞其意圖稱為對話行為(Dialogue act, DA)。最近由於深度學習的一大突破，許多在對話行為研究[6-9]都有比傳統機器學習方法有不錯的提升。對話行為分類在對話系統上佔據關鍵的因素，機器能夠充分了解使用者所表達的句子語意。

(一).系統介紹

1. 資料集簡介：本論文使用之資料集為去識別化後文字及簡訊客服(不含語音客服)的用戶對話資料，內容為行動電話與上網業務之常見客戶提問，擷取了其中較常見的 10 種提問意圖類別，並對語句進行下述文字正規化：(1)除去意圖不明語句(2)除去總字數小於 5 個字元內的語句(3)除去標點符號與未知字元(4)全形轉成半形、英文大寫轉成小寫。

   正規化後，共 15,773 筆訓練語句，1,584 筆測試語句，平均每句字數 12.61 個字，詳細說明如表一：

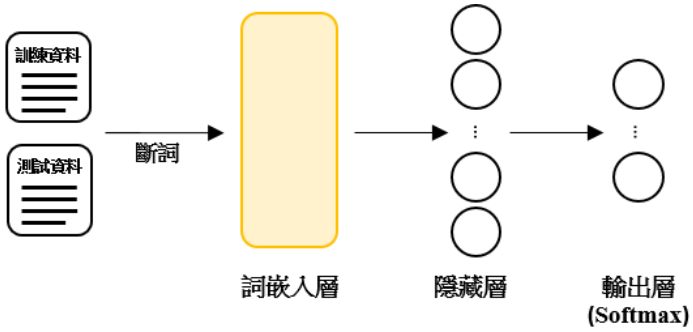
表一、資料集各類意圖筆數

意圖類別名稱	訓練集資料筆數	測試集資料筆數
資費與合約查詢或修改	2,000	253
帳單或繳費問題	2,000	102
加值服務問題	483	24
優惠方案	2,000	668
手機銷售問題	1,239	58
國際漫遊	2,000	113
手機用量問題	1,557	80
障礙申告或收訊問題	2,000	161
APP 使用問題	1,712	81
服務據點問題	782	44

(二).模型設計

近年來，人工智慧發展迅速，其中機器學習分支之一的深度學習在各個領域有許多重大突破。自然語言處理領域中，傳統機器學習方法需要人工設計模型所需的特徵組合，常消耗大量人力與時間。另一方面，深度學習則可以自動找出模型特徵表示，同時深度學習許多架構都能有效處理不同的自然語言處理任務，以下將介紹常見的深度學習模型。本論文討論以下 4 種預訓練詞向量方法對意圖分類器之影響：(1)隨機嵌入(2)Skip-Gram (3)BERT (4)BERT-WWM-Chinese，分類器模型訓練流程如圖

一：



圖一、分類器模型訓練流程圖

(1) 隨機嵌入：訓練分類器時，使用均勻分布(Uniform Distribution)隨機初始化詞嵌入層，每一詞彙索引對應一個值域範圍為[-0.05, 0.05]的 300 維詞向量。

(2) Skip-Gram[10]：Word2vec 是由 Mikolov 等人在 2013 年提出，使用到淺層的神經網路模型，模型分為 Continuous Bag of Words(CBOW)與 Skip-gram，CBOW 利用詞語的前後字建立詞窗當作輸入，來預測出此詞語，而 Skip-gram 則相反過來。由 Word2vec 產生的文字向量，將此向量投射到一個向量空間中，語意相近的詞彙將會在向量空間中非常相近，顯示文字可以在向量空間中有語意近似的關係。近期，許多的 Word2vec 應用在預訓練的詞向量作為訓練深度學習模型的詞語初始向量，對訓練模型時可以更有效的調整向量，加速收斂，同時也會不斷調整(fine-tune)詞向量，更能強化詞語語意關係。使用兩種資料集，分別為 2019 年 6 月繁體中文維基百科[1]，共 340,369 篇文章，斷詞[2]後共 1,042,225 個不重複詞彙；真實中文客服之對話資料集斷詞後共 1,653 個不重複詞彙，各自訓練兩個 300 維的詞向量模型。

(3) BERT(Bidirectional Encoder Representation from Transformers) [11]：從 ELMo[12]之後，根據前後文語意產生句向量之語意表達模型(Contextualized word representations)，成為近年來，自然語言處理領域的熱門研究主題，這類模型的優勢在於：(1) 能夠學習到詞彙在多情境下之不同語意、語法 (2) 能夠學習到詞彙依據上下文變化所帶來之不同語意，ELMo 透過雙向語言模型(Bidirectional language model, biLM)提供了對下游任務效果更佳的詞向量。而 OpenAI GPT[13]則將多層的

[1] https://dumps.wikimedia.org/zhwiki/
[2] https://taku910.github.io/crfpp/

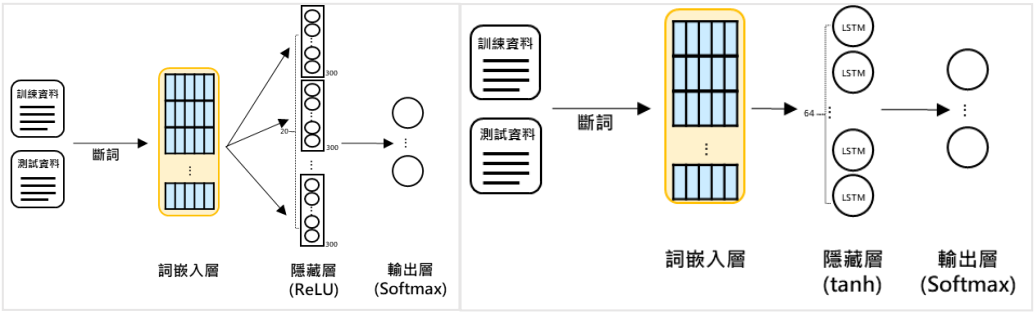
單向 Transformer[14]模組，引入了語意表達模型中。

Google 於 2018 年 10 月所發布的 BERT 模型，也基於 Transformer 模組，不同於 OpenAI GPT 模型的單向 Transformer，BERT 藉由從克漏字(Cloze)任務所帶來的靈感，使用雙向 Transformer 對遮蔽的 Token 進行預測(Masked Language Model, MLM)，以及預測下一句的下游任務，讓模型學習到句子之間的關係，建構一通用的預訓練語言表達模型(Language representation model)，使得自然語言處理的各項任務(如: 意圖分類、問答、翻譯等)，能夠使用較為輕量化的模型，輕易地進行遷移學習(Transfer learning)，在下游任務的訓練過程中，對表達模型進行優化(Fine tuning)即可。使用 Google 於 2018 年 11 月所發布的中文預訓練模型檔案，對每句訓練語句，產生 768 維的雙向語意表示(Bidirectional contextual representation)向量。

(4) BERT-WWM(Whole word masking)-Chinese[15]：中國哈爾濱工業大學於 2019 年 6 月發布，基於 2019 年 5 月 Google 所更新的 BERT 模型，新模型修改了原先訓練過程中，由隨機遮蔽單個字元，改為遮蔽完整詞彙(Whole word masking, WWM)，WWM 將完整詞彙的語意帶入模型中，使得模型較容易學習到字元間常用的組成關係。BERT-WWM-Chinese 採用與(3)相同的 BERT 模型作為基礎，使用中文維基百科及哈爾濱工業大學語言技術平台(Language technology platform, LTP)[3]的斷詞器，斷詞後，加入 WWM 重新訓練，產生 768 維的雙向語意表示向量。

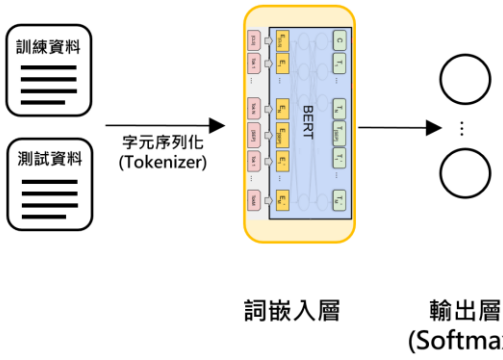
接續在(1)隨機嵌入與(2)Skip-Gram 兩種預訓練詞向量之後，本論文設計了兩種分類器，兩種分類器皆為多層感知機(Multi-Layer Perceptron, MLP)，包含一輸入層、一隱藏層(Hidden Layer)、一輸出層。兩種分類器的差異在於隱藏層之設計：對於句子中的斷詞結果，依每一詞彙出現的順序，產生對應的詞向量，如圖二，(a)將前 20 個詞彙的詞向量拼接起來，形成一 6000 維的句向量，連結一 64 維使用線性整流(Rectified linear unit, ReLU)函數作為激活函數的隱藏層。(b)在限制最多 20 個詞彙輸入的情況下，將詞向量依序放入，傳播至隱藏層共 64 維的長短期記憶(Long short-term memory, LSTM) [16]神經元。輸出為一 10 維隱藏層，使用 Softmax 函數作為激勵函數(Activation function)輸出該語句所預測之意圖分類。

[3] https://ltp-cloud.com/



圖二、詞嵌入層搭配 MLP/LSTM 之分類器

而在(3)BERT 及(4)BERT-WWM-Chinese 的雙向語意向量後，直接連結上述所提相同結構之輸出層，作為輸出該語句所預測之意圖分類，如圖三。



圖三、使用 BERT 及 BERT-WWM-Chinese 之分類器

三、結果與結論

(一). 實驗配置

使用 gensim[4]套件訓練 Skip-Gram 預訓練詞向量模型，過濾 5 個字元以下的語句，循環 10 次(Epochs)。而隨機嵌入與 Skip-Gram 所連接的 MLP、LSTM 分類模型，採 RMSprop 優化器(Optimizer)，學習率(Learning Rate)為 0.001，循環 50 次，批量大小(Batch size)為 64。BERT 與 BERT-WMM 使用 Adam 優化器，採 2e-5 的學習率(Learning rate)，循環 100 次，批量大小為 16。損失函數(Loss function)皆使用分類交叉熵(Categorical cross-entropy)，驗證(Validation)資料集大小皆設定為訓練資料集的 1%，並設定若連續 5 次模型在驗證集上的損失沒有下降，則提早終止(Early stopping)訓練過程。

[4] https://radimrehurek.com/gensim/models/word2vec.html

(二). 實驗結果

由於測試資料中，資料數量最多(668 筆)與最少(24 筆)的類別有不小之差距。為避免因為資料類別不平衡，導致實驗結果陷入正確率悖論(Accuracy paradox)，故採用分類正確率(Accuracy)之外，也使用了 Macro-F1 作為多類別意圖分類模型的評估指標。為了評估的一致性，每個模型皆隨機重覆跑了 5 次結果，取平均值及標準差，實驗結果如表二。

表二、比較不同詞嵌入層之分類器的結果

模型名稱			平均正確率(標準差)		平均 Marco-F1(標準差)	
			不更新詞嵌入層	更新詞嵌入層	不更新詞嵌入層	更新詞嵌入層
隨機嵌入		MLP	0.6830(0.026)	0.7456(0.019)	0.6920(0.026)	0.7290(0.017)
		LSTM	0.6506(0.030)	0.7322(0.046)	0.5956(0.022)	0.6782(0.020)
Skip-Gram	維基百科	MLP	0.6988(0.016)	0.7664(0.008)	0.6718(0.007)	0.7516(0.006)
		LSTM	0.7792(0.016)	0.8106(0.016)	0.7652(0.015)	0.8044(0.014)
	客服資料	MLP	0.6978(0.016)	0.7838(0.010)	0.6656(0.005)	0.7672(0.009)
		LSTM	0.7862(0.020)	0.8110(0.013)	0.7638(0.013)	0.8040(0.013)
BERT				0.8290(0.022)		0.8300(0.015)
BERT-WWM-Chinese				**0.8414(0.006)**		0.8435(0.017)

從表二的實驗數據中，可以得到以下資訊：

1. BERT 架構正確率較高：使用 BERT 架構的模型，與其他模型相比，於分類正確率與 Macro-F1 分數皆取得較好的成績；導入 WWM 之後，與原 BERT 模型相比，兩種指標也有些微提升。

2. 更新詞嵌入層優於不更新：對於每一種預訓練詞向量模型，在訓練分類器的過程持續更新詞嵌入層，分類正確率進步至少 3%左右。

3. MLP 與 LSTM 之比較：值得注意的是，使用隨機嵌入，MLP 會得到比 LSTM 更好的效果，但使用 Skip-Gram 作為預訓練詞向量則反之。

4. 隨機嵌入法與 Skip-Gram 之比較：同是使用客服資料進行訓練的隨機嵌入法與

Skip-Gram，不管使用 MLP 或 LSTM 作為分類器，Skip-Gram 的效果都比隨機嵌入法來的佳。

5. 維基百科及客服資料之比較：兩者用來訓練 Skip-Gram，發現在兩種指標上皆沒有太大的差異，但兩者的語料規模差異很大；顯示若使用與下游任務領域相符的語料，訓練詞向量模型，較可節省計算資源，達到使用一般詞彙語料訓練之水平。

使用維基百科訓練的 Skip-Gram 之 LSTM 分類器，有 70%的錯誤分類來自「優惠方案」，22%的錯誤分類來自「資費與合約查詢或修改」；而使用 BERT 訓練之分類器，有 75%的錯誤分類來自「資費與合約查詢或修改」11%的錯誤分類來自「優惠方案」。這兩個意圖類別也是資料筆數最多的類別，其中可以發現收集到的語句資料變異情況較大，且單詞容易與其他意圖類別之語句重複，故容易分類錯誤。

四、結論

近年來詞嵌入向量(word embedding)在自然語言研究中激起一股研究熱潮，此種表示方式，不僅能以較低維度的向量表示詞彙，還能藉由詞向量間的運算，找出任兩詞彙之間的語意關係。因此，本文實驗探討了新穎的詞向量模型 BERT 與隨機嵌入/Word2Vec 對於特定領域之客服對話意圖辨識效能增進的議題。主要貢獻有兩個部分: 第一部分，本論文將詞向量表示資訊應用於特定領域的客服系統中，透過此種表示方式而能獲取到更多詞彙間的語意資訊，以提升意圖辨識的準確度。第二部分，我們對於近來深度學習強調使用預訓練模型進行微調的訓練模式進行了驗證。我們利用 BERT 模型以及其增益模型 BERT-WWM-Chinese 為預訓練模型，使用少量的訓練資料進行微調，確實幫助意圖辨識準確率的提升。未來，勢必會有更多新穎的預訓練模型不斷的被提出來加強各種 NLP 後端應用的效能。因此，能有效評價各模型於不同後端應用的優劣與整合不同模型達到集成學習的效果也是未來研究重點之一。

參考文獻

[1] S. J. Young, "Probabilistic methods in spoken–dialogue systems," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* vol. 358, no. 1769, pp. 1389-1402, 2000.

[2] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv preprint arXiv:1605.07683*, 2016.

[3] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models," in *AAAI*, 2016, vol. 16, pp. 3776-3784.

[4] T.-H. Wen *et al.*, "A network-based end-to-end trainable task-oriented dialogue system," *arXiv preprint arXiv:1604.04562*, 2016.

[5] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," *arXiv preprint arXiv:1703.01008*, 2017.

[6] L. Meng and M. Huang, "Dialogue Intent Classification with Long Short-Term Memory Networks," Cham, 2018: Springer International Publishing, in Natural Language Processing and Chinese Computing, pp. 42-50.

[7] H. Kumar, A. Agarwal, R. Dasgupta, and S. Joshi, "Dialogue act sequence labeling using hierarchical encoder with crf," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[8] C. Cerisara, P. Kral, and L. Lenc, "On the effects of using word2vec representations in neural networks for dialogue act recognition," *Computer Speech & Language*, vol. 47, pp. 175-193, 2018.

[9] S.-s. Shen and H.-y. Lee, "Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection," *arXiv preprint arXiv:1604.00077*, 2016.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] M. E. Peters *et al.*, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[13] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[14] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[15] Y. Cui *et al.*, "Pre-Training with Whole Word Masking for Chinese BERT," *arXiv preprint arXiv:1906.08101*, 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.