使用對話行為嵌入改善對話系統用戶訊息中提問句與閒聊句之判別

Improve Chit-Chat and QA Sentence Classification in User Messages of Dialogue System using Dialogue Act Embedding

Chi-Hsiang Chao
ChingShin Academy
Taipei, Taiwan
10835028@st.chjhs.tp.edu.tw

Xi-Jie Hou
ChingShin Academy
Taipei, Taiwan
10935020@st.chjhs.tp.edu.tw

Yu-Ching Chiu
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
crystalchiu@g.ncu.edu.tw

摘要

近年來，對話系統蓬勃發展並被廣泛應用於客服系統並取得了不錯的成效。檢視用戶與真人客服間的對話紀錄，可以發覺用戶的語句夾雜著對產品與服務的問題，以及和客服之間的閒聊。根據專業人員的經驗，在客服對話中適當地夾雜閒聊有助於提升用戶的體驗。然而，用戶提問是期望獲得解答，閒聊則是期望與客服有人與人之間的互動交流。面對這兩種意圖，對話系統必須能有效判別，以產生適當的回應。對話行為 (Dialog Act) 是語言學家將對話語句依據其作用定義出的一種分類方式。我們認為這個資訊將有助於提問句及閒聊句的區分。在本研究中，我們結合一個已公開的Covid-19問答資料集及一個Covid-19主題的閒聊資料集組成我們的實驗資料。我們基於 BERT (Bidirectional Encoder Representation from Transformers) 模型建立了一個提問句—閒聊句分類器模型。實驗結果顯示，加入對話行為嵌入 (Dialog Act Embedding) 的組態比僅使用原始語句嵌入的組態準確率高了16%。此外，經過分析發現，Statement-non-opinion、Signal-non-understanding、Appreciation 等對話行為類型與提問句較相關，Wh-Question、Yes-No-Question、Rhetorical-Question 等類型則與閒聊句較相關。

Abstract

In recent years, dialogue system is booming and widely used in customer service system, and has achieved good results. Viewing the conversation records between users and real customer service, we can see that the user's sentences are mixed with questions about products and services, and chat with customer service. According to the experience of professionals, it is helpful in improving the user experience to mix some chats in customer service conversations. However, users' questions are expected to be answered, while chatting is expected to interact with customer service. In order to produce an appropriate response, the dialogue system must be able to distinguish these two intentions effectively. Dialog act is a classification that linguists define according to its function. We think this information will help distinguishing questioning sentences and chatting sentences. In this paper, we combine a published COVID-19 QA dataset and a COVID-19-topic chat dataset to form our experimental data. Based on the BERT (Bidirectional Encoder Representation from Transformers) model, we build a question-chat classifier model. The experimental results show that the accuracy of the configuration with dialog act embedding is 16% higher than that with only original statement embedding. In addition, it is found that conversation behavior types such as "Statement-non-opinion", "Signal-non-understanding" and "Appreciation" are more related to question sentences, while "Wh-Question", "Yes-No-Question" and "Rhetorical-Question" questions are more related to chat sentences.

關鍵字：對話行為、對話系統
Keywords : Dialog act classification, Dialog system

1 緒論

現今社會的客服系統多以對話系統回應用戶端，不僅能節約成本，更能在同一時間內解決大量的問題。在現實生活中，用戶與客服系統的對話中時常夾雜著「提問」和「閒聊」的語句，而這樣的對話方式有助於提升用戶的體驗。然而，提問是期望獲得解答，而閒聊則是希望與對方有所交流。因此有效的判別用戶的意圖，對於對話系統產生適當的回應十分重要。現代的對話系統大多是單一功能的系統，如：任務導向式對話系統 (Task-Oriented)、閒聊式對話系統 (Chit-Chat) 和問答

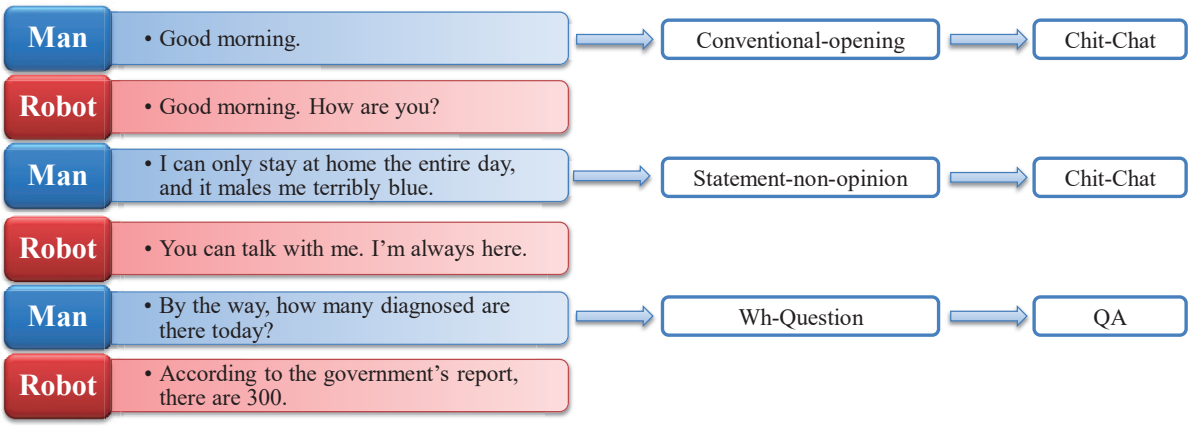

圖 1. 使用者與機器人的對話

式對話系統 (Question Answering) 等。例如：Google 助理以及 Siri 較偏向於任務式導向式對話系統，在以對話了解用戶需求後為用戶執行任務；閒聊 (Chit-Chat) 則是允許用戶與系統進行開放式聊天對話，Google 的 Meena (Adiwardana, et al., 2020)即屬之；問答式 (Question Answering) 是接收用戶端所提出的問題，從資料庫中尋找最佳的答案，再回答用戶端，IBM 的 Watson (Gliozzo, et al., 2013)即屬之。以上幾種對話系統雖然都能專精於特定領域，能提供用戶端正確的回應，但是目前對話系統的功能較為單一化，無法集上述三種對話方式於一身，使得對話系統較難運用在真實的對話場景中。為使聊天機器人通用於上述場域中，本研究引入對話行為 (Dialogue Act) (Perkoff, E Margaret, 2021)，發展辨識用戶話語意圖究竟為發問或閒聊的技術。對話行為是語言學家將對話的句子依據其作用及定義得出之分類，我們將對話行為作為分辨「問答式對話」和「閒聊式對話」的重要參考資訊，使對話系統模型能將用戶的話語導向問答模組或閒聊模組以給予回應，使兩種對話系統能合而為一，以提供更好的用戶體驗。本研究將探討是否能夠利用機器學習模型有效地藉由引入對話行為區分出問答和閒聊兩種情境的語句，以便後續利用對應的系統進行後續處理與應對。

圖 1 為使用者與對話系統之間的對話，可以看出使用者的前兩句話屬於閒聊 (Chit-Chat)，最後一句則是提問 (QA 中的 Question)。本研究探討對話行為對於模型判斷句子屬於 Chit-Chat 或 QA 是否有幫助，如同圖中的對話系統由 Conventional-opening 和 Statement-non-opinion 判斷屬於 Chit-Chat、由 Wh-Question 判斷屬於 QA 一般。

2 相關研究

2.1 Hybrid Dialogue System

大部分任務型導向對話系統與閒聊型對話系統通常單獨出現，儘管現今已有許多針對上述兩種對話系統的研究 (Hosseini-Asl, et al., 2020) (Adiwardana, et al., 2020)，然而混和這兩種對話系統的模型尚未有足夠的研究。此論文(Moirangthem, Dennis Singh, et al., 2018)試圖以將任務導向的句子與閒聊的句子區分出來的方式，構建一個混合任務型導向對話系統與閒聊型對話系統的模型。

對話系統大多無法兼具多種功能。此篇論文 (Sun, Kai, et al., 2020) 嘗試通過添加較隨意且與上下文相關的閒聊語料增強任務導向聊天機器人的對話能力，並綜合這兩種類型的系統。目的是讓一個虛擬助理能使任務式與閒聊型的機器人合二為一，加強在兩者之間切換的能力，以及強化系統對話的趣味以及交流性，使之更像人類，藉此提升用戶的使用體驗。綜上，混合型對話系統將成為未來對話系統發展的趨勢。

2.2 Dialogue Act in Dialogue system

對話行為 (Dialogue Act)是一種從句子中抽取出來的語意抽象，表示句子在對話中的功能，即此句子背後的行為。例如：" Hi. How are you? " 的對話行為是" Greeting "，因為此句話

代表的動作為「打招呼」。為了增加句子所包含的資訊，(Kumar et al., 2018) 將對話行為加入到句子中來訓練對話系統，並使其預測給使用者的回應 (next utterance selection)，對話行為反映出了對話系統與用戶之間的對話模式，對話行為提供的資訊有效改善了系統回應的表現。

在本研究中，我們將探討加入對話行為對判斷用戶語句意圖為提問或閒聊所帶來的助益，並討論不同種加入對話行為的方式對模型表現的影響。

3 方法

3.1 BERT

本文選擇訓練 BERT (Devlin, Jacob, et al., 2019) (全名為 Bidirectional Encoder Representations from Transformers) 區分 QA 及 Chit-Chat。BERT 是 Google 以無監督的方法利用大量無標註文本訓練的語言代表模型，其架構為 Transformer (Vaswani, et al., 2017)中的 Encoder。相較於其他的語言代表模型，BERT 以漏字填空(MLM)和下個句子預測 (NSP)的兩個任務訓練出一個能被廣泛用於理解自然語言的模型，再訓練此模型做 Fine-tuning 的監督式任務，來達到其目的。BERT 主要可以做四個下游任務模型，分別是單一句子分類任務、單一句子標註任務、成堆句子任務、問答任務，而本次實驗所使用的是單一句子分類任務。

3.2 DialogTag

DialogTag 是 Bhavitvya Malik 基於 38 種對話行為所製成的分類器，它能夠將輸入句子所屬的對話行為輸出。它是使用 Tensorflow 建立的一個 Transformer 模型，在 Python 中有釋出專門的模組供大家使用。它的原型是賓州大學發表的 Switchboard Corpus (雙向通話紀錄) (Godfrey, et al., 1992)，其中共分類出 42 種對話行為，DialogTag 從中抽取 38 種對話行為作為簡化過後的版本。

本研究參考了(Stolcke, Andreas, et al., 2000) 的對話行為，探討問答語句及聊天語句是否能透過對話行為進行準確的區分。我們將資料輸入 Bhavitvya Malik 所釋出的對話行為系統分類器 (DialogTag) 並分析結果，最後確認是否可以對話行為區分出問答和閒聊兩種情境。我們將句子和對話行為的嵌入表示連接成一個新的句子嵌入表示作為輸入，訓練以 BERT 為基底的二元分類模型，預測當前輸入的句子為 Chit-Chat 或是 QA，模型架構如圖 2 所示。

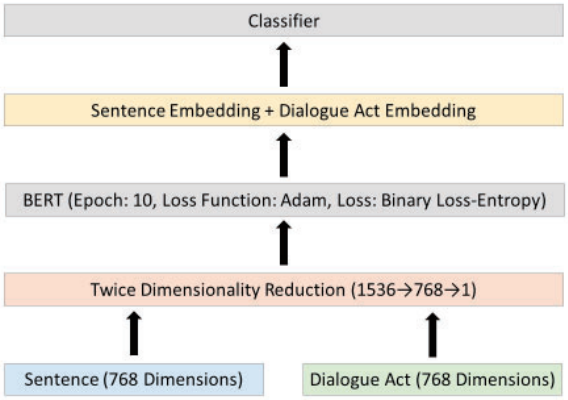

圖 2. 模型架構

4 資料集

4.1 SQuAD2.0

SQuAD2.0 (Rajpurkar, Pranav, et al., 2018) 是由超過 50000 個無法回答的問題和 2016 年 Rajpurkar 等人所發布的斯坦福問答數據集 (SQuAD (Rajpurkar et al., 2016)) 所組成的，每個問答對都有一個給定的上下文段落，成為閱讀理解任務的常用測試資料集。我們選擇 SQuAD2.0 是因為 2018 年 OpenAI-GPT 和 BERT 在使用 SQuAD2.0 做多語言任務上取得了很好的效能。它將作為問答用的資料。

4.2 Dataset for chatbot Simple questions and answers

我們使用由 Graf Stor 在 Kaggle 所釋出的 Dataset for chatbot Simple questions and answers[1]作為閒聊的資料集。此資料集是作者本人為了訓練簡單的 Seq2Seq (Sutskever, Ilya, et al., 2014) 而蒐集而成的。

[1] Stor, Graf. "Dataset for Chatbot." *Kaggle*, 14 June 2020, www.kaggle.com/grafstor/simple-dialogs-for-chatbot.

4.3 COVID-QA

我們使用由 Xing Han Lu 所釋出的 COVID-QA[2] 資料集，為關於 COVID-19 (與新聞、公共衛生和社區討論的關係較為密切) 的問答句子。此資料集是為了方便提供問答系統的建立而蒐集的 1800 多筆問答資料，因此我們將其當作問答句的資料。

4.4 COVID-19

我們所使用的是由 Moayad 釋出的 COVID-19：Audience-LiveChat[3] 資料集。此資料集為有關 Covid-19 的 YouTube 影片的直播聊天室內容，共有大約 73 萬筆閒聊對話。此為閒聊資料集。

4.5 Switchboard Corpus

我們將 Switchboard Corpus 作為另一個閒聊資料的來源。它擁有大約 2400 組雙向對話紀錄與接近 260 個小時的總對話時數。

5 實驗

5.1 訓練資料、驗證資料和預測資料

這次實驗總共做了 6 個模型以判斷對話行為是否能有助於模型分辨 QA 和 Chit-Chat：BERT_General 、 BERT_General_DAC 、 BERT_General_Concatenation 、 BERT_COVID 、 BERT_COVID_DAC 與 BERT_COVID_Concatenation 。 其 中 的「COVID」是在訓練的過程中加入較雜亂的 Chit-Chat 句子，例如「？？？？？」和「omg」，以增強模型對於 Chit-Chat 句的判斷能力。在訓練完 BERT_General 後，我們發現在 BERT_General 的輸出結果中，答錯的句子多為這類混亂的 Chit-Chat 句子，可見此模型應付這類句子的能力較低，因此我們才加入它們以提升模型的準確率。「DAC」與「Concatenation」則是將句子串接對話行為，兩者之間的差異在於串接的方式。「DAC」先將這兩部分分別轉為 e 向量後再直接連接起來，合併為一個 1536 維的向量，再送入分類器進行 Chit-Chat 與 QA 的分類。「Concatenation」則是一開始就將句子與對話行為的文字相連，再一起轉為 768 維向量並送入分類器。

表 1 為各個模型所使用的資料數量。資料來如下：(1) Chit-Chat 所用的資料集來自 Dataset for chatbot Simple questions and answers 以及 Switchboard Corpus；(2) QA 所用的資料集來自 SQuAD2.0；(3) COVID-19 相關的 Chit-Chat 資料來自 COVID-19： Audience-LiveChat；(4) COVID-19 有關的 QA 資料來自 COVID-QA。在建立三種資料集時，維持 Chit-Chat 和 QA 的數量比為 1：1，訓練資料和驗證資料的數量比為 8：2。我們最終用來測試模型的預測資料總共有 1900 筆，含有 950 筆 Chit-Chat（全部與 COVID-19 相關）以及 950 筆 QA（全部與 COVID-19 相關）。訓練資料和驗證資料所使用的資料筆數如表 1 所示，表 2 為 22 種出現在資料中的對話行為。

	Train Data			Valid Data		
	# Chit-Chat	# QA	Sum	# Chit-Chat	# QA	Sum
BERT_General	10500	10500	21000	2625	2625	5250
BERT_General_DAC	10500	10500	21000	2625	2625	5250
BERT_General_Concatenation	10500	10500	21000	2625	2625	5250
BERT_COVID	10500 (with COVID data)	10500	21000	2625 (with COVID data)	2625	5250
BERT_COVID_DAC	10500 (with COVID data)	10500	21000	2625 (with COVID data)	2625	5250
BERT_COVID_Concatenation	10500 (with COVID data)	10500	21000	2625 (with COVID data)	2625	5250

表 1. 各模型使用的資料量

Acknowledge (Backchannel), Action-directive, Appreciation, Collaborative, Conventional-closing, Conventional-opening, Declarative Yes-No-Question, Hold before Answer/Agreement, Negative, Non-no Answers, No Answer, Open-Question, Or-Clause, Other, Quotation, Repeat, Rhetorical-Question, Self-talk, Signal-non-understanding, Statement-non-opinion, Statementopinion, Wh-Question, Yes-No-Question

表 2. 出現在資料中的對話行為，共 22 種

5.2 實驗與結果

表 3 為各模型在分類 Chit-Chat 及 QA 的表現與整體的準確率，分為以下層面討論: (1)不

[2] Xhlulu. “COVID-QA.” *Kaggle*, 15 Apr. 2020, www.kaggle.com/xhlulu/covidqa.

[3] Moayad. “COVID-19: Audience-LiveChat.” Kaggle, 17 Apr. 2020, www.kaggle.com/moayadhn/covid19-roylablivechat.

同訓練資料的差異 (2) 有無加入對話行為的差異(3)不同加入對話行為方法的差異。

	QA Accuracy	Chit-Chat Accuracy	Total Accuracy
BERT_General	61.4%	45.3%	58.1%
BERT_General_DAC	76%	61.4%	68.7%
BERT_General_Concatenation	80.7%	78%	79.4%
BERT_COVID	52.5%	99.8%	76.2%
BERT_COVID_DAC	66%	99.5%	82.8%
BERT_COVID_Concatenation	85.4%	99%	92.2%

表 3. 模型準確率

5.3 不同訓練資料的差異

比較表 3 中 BERT_General 和 BERT_COVID 的數據可以看出加入較混亂的 Chit-Chat 資料大幅提升了模型判別 Chit-Chat 的能力，準確率由 45.3%上升到 99.8%，進步了 54.5%。雖然使判斷 QA 的準確率略微下降，但仍然使模型整體的準確率提升 29.4%。

5.4 有無加入對話行為的差異

由表 3 中 BERT_COVID 和 BERT_COVID_DAC 的數據比較可以看出在加入對話行為後，QA 的準確率會提升。在表中，BERT_ COVID 在 QA 上僅有 52.5%的準確率，而 BERT_ COVID_DAC 則達到了 66%的準確率。在整體表現上，BERT _COVID_DAC 亦達到了 82.7%的高準確率，相較 BERT_ COVID 有 6.6%的進步，由此可見加入對話行為確實能有效的提升整體的分類準確率。

5.5 不同加入對話行為方法的差異

從表 3 的數據中可以看出不同加入對話行為的方式確實會影響模型的判斷。結果顯示 BERT_General_concatenation 的準確率比 BERT_General_DAC 高出 10.7%。BERT_COVID_Concatenation 則是比 BERT_COVID_DAC 準確 9.5%。因此，將對話行為與句子以文字方式連接擁有全部模型中最高的準確率。

5.6 資料分析

對於模型所得出的資料，我們亦以 Visual Correlation 的方式進行分析。Visual Correlation 指的是在兩組資料之間找出關聯性，並採用視覺化的方式表現。Visual Correlation 更能將數據之間的差異展現出來，比較容易讓人找出相關性高的部分。基於以上原因。我們採用 Visual Correlation 分析資料。

5.7 視覺化

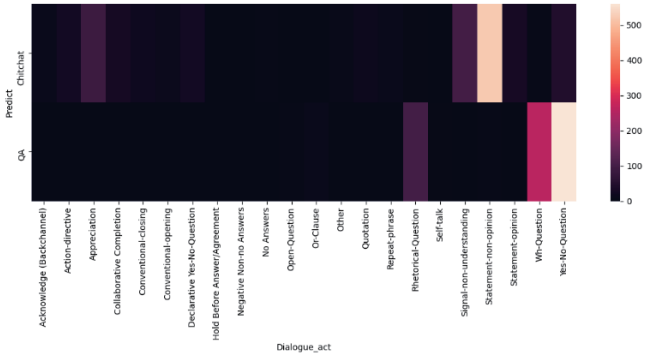

圖 3. BERT_COVID_Cocatenation 預測結果與對話行為的 Visual Correlation

Visual Correlation 指的是找出兩組數據之間的關聯性，再將其以視覺化的方式展現出來。Visual Correlation 更容易顯示出資料間的差異，也更容易找出關聯性高的項目。基於以上的原因，我們決定使用 Visual Correlation 來分析模型產出的資料。從以上的實驗，我們發現加入對話行為對於分辨閒聊與問答有所幫助。此外，我們將 Visual Correlation 套用在預測出來的資料上得到標籤（閒聊與問答）以及對話行為之間的熱圖以找出影響模型的對話行為。

圖 3 為 BERT_COVID_Cocatenation 預測結果與對話行為的 Visual Correlation，越淺色的部分相關性越高。從圖中可以發現，閒聊句與「Statement-non-opinion」高度相關，而問答句與「Yes-No-Question」和「Wh-Question」有關，其餘的對話行為則沒有顯著的相關性。因此，在以對話行為替句子分類時，若是遇到「Statement-non-opinion」，可以將句子送往閒聊機器人；反之如果是「Yes-No-Question」或「Wh-Question」則是輸入進問答機器人。

6 結論

本次研究使我們發現加入對話行為確實對於模型的準確率有所幫助。以 BERT_COVID 與 BERT_COVID_Concatenation（我們建立的模型）為例，後者比前者多了 16%的準確率，

同時也比其他所有模型都優秀，達到了92.2%的準確率。

除此之外，我們也找出三種可以有效區別閒聊與問答的對話行為，分別是「Statement-non-opinion」、「Yes-No-Question」與「Wh-Question」。有些對話行為，例如「Appreciation」和「Rhetorical-Question」，比上述幾個的關聯性都還要低，但它們也可以提升模型的分類能力。有了對話行為的幫助，輸入句子可以被送到對應的機器人已產生適當的回應。

7 未來展望

在本論文中，我們進行處理的都是單一句子，沒有考慮前後文，這限制了我們的研究。我們的模型誤將「Answer」的對話行為判定為閒聊，但實際上這應該屬於問答的一部份。要解決這方面的問題，後續的研究可以將對話紀錄納入考量。加入對話紀錄的模型能夠透過前文與當下的句子判斷此劇屬於閒聊或是問答，特別是問答句的準確率可以大幅提升，因為在知道先前的問題下，「Answer」就會被分類為問答句。

References

D. Adiwardana, M.T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu. 2020. *Towards a humanlike open domain chatbot.* arXiv preprint arXiv:2001.09977.

Alfio Gliozzo, Or Biran, Siddharth Patwardhan, Kathleen McKeown. 2013. *Semantic Technologies in IBM Watson™.* Proceedings of the Fourth Workshop on Teaching Natural Language Processing, pages 85–92

Perkoff, E Margaret. 2021. *Dialogue Act Analysis for Alternative and Augmentative Communication.* Proceedings of the 1st Workshop on NLP for Positive Impact, pages 107—114

Hosseini-Asl, Ehsan and McCann, Bryan and Wu, Chien-Sheng and Yavuz, Semih and Socher, Richard. 2020. *A simple language model for task-oriented dialogue.* arXiv preprint arXiv:2005.00796.

Dennis Singh Moirangthem and Minho Lee. 2018. *Chat Discrimination for Intelligent Conversational Agents with a Hybrid CNN-LMTGRU Network.* Proceedings of The Third Workshop on Representation Learning for NLP.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2020. *Adding Chit-Chats to Enhance Task-Oriented Dialogues.* arXiv preprint arXiv:2010.12757.

Harshit Kumar, Arvind Agarwal, Sachindra Joshi. 2018. *Dialogue-act-driven Conversation Model: An Experimental Study.* Proceedings of the 27th International Conference on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia. 2017. *Attention is all you need.* Advances in neural information processing systems, pages 5998--6008.

Godfrey, John J and Holliman, Edward C and McDaniel, Jane. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Acoustics, Speech, and Signal Processing, IEEE International Conference on. IEEE Computer Society, pages 517--520.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. *Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech.* Computational linguistics 26.3, pages 339--373.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. *Squad: 100,000+ questions for machine comprehension of text.* arXiv preprint arXiv:1606.05250.

Pranav Rajpurkar, Robin Jia,and Percy Liang. 2018. *Know What You Don't Know: Unanswerable Questions for SQuAD*. arXiv preprint arXiv:1806.03822.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. *Sequence to Sequence Learning with Neural Networks*. Advances in neural information processing systems.

Benjamin Kane, Georgiy Platonov, and Lenhart Schubert. 2020. *History-Aware Question Answering in a Blocks World Dialogue System*. arXiv preprint arXiv:2005.12501.