https://doi.org/10.2352/EI.2023.35.3.MOBMU-359

Performance Evaluation of Keyword Detection for the Chatbot Model

Ganesh Reddy Gunnam*[1] *, Devasena Inupakutika*[1] *, Rahul Mundlamuri*[1] *, Sahak Kaghyan, David Akopian; The University of Texas at San Antonio; Patricia Chalela, Amelie G. Ramirez; The University of Texas Health Science Center; San Antonio, Texas

Abstract

To enable enriched free-text human-computer conversations, keyword detection is an important component in chatbot models as it helps to identify specific keywords in user inputs that can trigger the chatbot to respond in a certain way. The performance of keyword detection thus depends on several factors such as the quality and quantity of training data, the selection of apt learning algorithms, and the tuning of various parameters. The performance evaluation of keyword detection for a chatbot involves taking into consideration, the specific requirements of the chatbot and the expected usage patterns of its users. To that end, this work investigates the keyword detection performance with limited vocabulary in the closed-domain chatbot model. A keyword reduction methodology is presented and experimental results on one public and one custom closed-domain datasets indicate about 4.6% improvement in F1-score and comparable performance respectively.

Introduction

The adoption of chatbots has been on the rise due to the increasing number of users adapting to conversations with their devices at the same level as with humans. By 2024, the global chatbot market is expected to reach 2 billion dollars [1]. Chatbots are generally text-based dialog agents that belong to the category of conversational systems and have invaded the world of instant messaging [2]. Chatbots are used in various applications such as helping us navigate complex technical issues, booking flight or hotel reservations, filing bank or insurance claims, and booking doctor appointments, among others. The naturalness of the chatbots determines their ability to facilitate high-quality and efficient, human-machine interfaces.

With regards to chatbots' scope and their generated responses, chatbots are either open-domain or closed-domain. Open-domain chatbots imitate human conversations on a wide range of topics. Thus, they are typically used for entertainment, marketing, and socializing. Earlier open-domain chatbots such as ELIZA [2] are the first examples of text-based dialog agents. Other remarkable chatbots are A.L.I.C.E (developed using Artificial Intelligence Markup Language) [3], Cleverbot [4] and Jabberwacky [5], Mitsuku [6], among others. A significant number of handcrafted rules and structured question-answers are used to build such chatbots.

Human conversations are much more complex than task-oriented chatbots can handle with simple handcrafted rules and templates. Natural language is not always clear and structured, and humans often use slang, colloquialisms, and expressions that are difficult for chatbots to understand. Additionally, humans often change the topic of conversation or ask unexpected questions (out of vocabulary), which can be challenging for chatbots to respond to without context [7] [8]. While successful conversations require understanding the keywords correctly, keyword annotation can be ambiguous. The users' utterances may contain certain words or free text, causing multiple keywords to be present in the same utterance. This keyword ambiguity is challenging for chatbots.

To address this, chatbots can use machine learning algorithms to improve the accuracy of keyword labeling. These algorithms can learn from a large corpus of human conversations and can recognize patterns and context in the user's utterances. Additionally, chatbots can use entity recognition to identify important information in the user's utterance, such as dates, locations, and names, to provide more accurate and relevant responses. To meet user preferences, open-domain chatbots need a rich set of keywords (vocabulary) to train the chatbot model. However, closed-domain chatbots are usually task-oriented. Thus, the chatbot model needs limited keywords to be detected. This paper studies the impact of limited vocabulary (number of keywords) on keyword detection performance utilizing the strength of a directed dialog in rule-based chatbots (pre-defined keywords) with the flexibility of free-text (natural language). The case-study chatbot consists of predefined rules with conversational capabilities for promoting smoking cessation [9].

The remainder of this paper is organized as follows. Section II presents related work and chatbot preliminaries. Section III covers keyword detection methodology with dataset details and model architecture. Section IV presents experimental results are discussed. Conclusion and future work are present in section V.

Related Work and Chatbot Preliminaries

The remarkable demand and interest in chatbots and their applications in the recent years highlighted the need in artificial-intelligence (AI) based chatbot applications with NLP (natural language processing) [10] support for human-like conversations. Such chatbots are mostly end-to-end probabilistic and neural conversational models-based [2], and data-driven systems. Recently, chatbots have been able to direct users in an appropriate direction while being able to cover a broad spectrum of user inputs. Bots such as Xiaoice [11], Zo (Microsoft), and Meena [12] aim to achieve human-like conversations with large, free-form datasets [2] through further adaptable and scalable learning processes.

On the contrary, the primary purpose of closed-domain, or task-oriented chatbots [2] is to assist users in performing specific tasks, such as customer support, primary care, out-patient healthcare management, coaching or counselling tasks, and healthcare

[1]These authors contributed equally.

interventions [13]. Such chatbots automate existing routine tasks and human-centered processes, allowing organizations to directly interact with their customer base. These chatbots also have added benefits such as always-available services, improved customer experiences, and reduced costs that are typical requirements of industries like healthcare and finance. Nevertheless, users input new and unexpected [14] scenarios as free-text, demanding a human facilitator when required. These scenarios are processed differently by closed- or open-domain chatbots. For open-domain chatbots, such conversations may sometimes get stuck in a loop until a point is reached where the chatbot recognizes the topic inside its range [15]. Closed-domain chatbots are constrained to domain-specific vocabulary and keywords. Therefore, outside-scope conversations tend to be directed to a default keyword, require a human interface, or end abruptly if not handled properly. The closed-domain chatbots with simpler domain models well understood by users like the case of transactional applications (ordering pizza, flight reservations or status, and banking services, among others) can be realized via directed dialog. Such applications follow a structured conversation approach. Despite being restrictive and having built on constrained vocabulary, directed dialog can have a much higher usability and improvement in text recognition rates [16]. Moreover, with directed conversation strategy, users will not be lost, and would be aware of system's expected input and capabilities.

Therefore, in this work, we focus on keyword detection for the closed-domain chatbot model with a limited vocabulary to be identified using the CBOW (continuous bag of words) [17] model.

Keyword detection methodology

This section presents the approach applied to selectively remove keywords from the full vocabulary dataset being used for training. Keywords should be distinct to help distinguish examples, which in turn enables interpreting the content of a corpus effectively. The topic-based vocabular size studies in [18] assert that lower number of keywords contribute to lower information entropy. While the overall vocabulary size does not play a big role. The keyword count additionally affect the pairwise topic similarity (PTS). PTS reveals how distinct the trained keywords are. The number of keywords with the full vocabulary favors larger number of keywords for higher distinctiveness.

The influence of keyword count on PTS exists due to the presence of most frequently occurring words across the examples. To that end, in order to reduce the number of keywords, the c-tf-idf matrix of examples is calculated and reduced by iteratively merging the least frequent keyword with the most similar one based on their c-tf-idf matrices. The keywords, their representations, and their frequencies/ sizes are updated eventually. This updated dataset is utilized for keyword detection. The performance results with the varying keyword count on 1 public dataset (Movie Tag: 100, 50, and 40) and 1 custom closed-domain chatbot corpus datasets (28, 14, and 7) are evaluated in this work.

A total of 3 keyword counts such as 7, 14, and 28 for custom chatbot corpus are compared based on two of the metrics that quantify keyword characteristics namely, pairwise topic similarity (PTS) and information entropy. PTS between 2 keywords is measured as the cosine value of their keyword vectors. IE of each example is averaged across examples in the collection. It is to be noted that we keep the total examples in the dataset same across the keyword counts. Figure 1 shows the low entropy and high PTS values for the lower number of keywords.

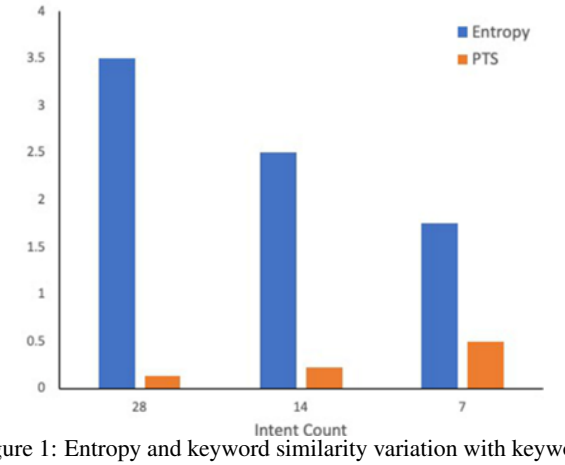

Figure 1: Entropy and keyword similarity variation with keyword count

Dataset details and Statistics

This work performs evaluation of the keyword detection methodology with varying keywords count on the following two datasets: a) public Movie Tags dataset [19], and b) custom smoking cessation chatbot corpus (section IV.A). The details of both the datasets are present in table 1.

Table 1: Dataset Statistics for Keyword Prediction.

Dataset	Keyword count	Chatbot Style	Training data
Movie Tags (Plot synopsis)	100	No	Constrained
Case study Chatbot Corpus	28	Yes	Constrained

Movie Tags Dataset: The Movie Tag dataset is a corpus of movie plot synopses and tags. The dataset consists of 4172 movie plots that are taken from IMDB and 10656 movie plots that are taken from Wikipedia. The dataset contains all the IMDB id, title, plot synopsis, tags for the movies. Overall there are about 14828 movies' data. Table 2 consists of some examples from the dataset.

Custom Chatbot Corpus: Keyword detection is a crucial part of task-oriented conversational systems. However, there are not many public keyword detection datasets available for testing and evaluation of such specific chatbot models. We built the custom chatbot dataset that is a short-query dataset with about 28 keywords. This custom dataset is created utilizing a Smoking Cessation-based healthcare intervention protocol by parsing the 6-month long short-queries outlined by UTHSCSA Public Health research team [9] and consists of 200 questions, a total of 7200 examples and 28 different keywords that were extracted from the manually annotated conversations. Some example keywords are shown in table 3.

Table 2: Dataset details: Movie Tags Dataset.

Imdb_Id	Title	Plot_Synopsis	Tags	Source
tt0057603	I tre volti della paura	Note: this synopsis is for the original Italian release with the segments in this certain order...	cult, horror, gothic, murder, atmospheric	imdb
tt1733125	Dungeons and Dragons: The Book of Vile Darkness	Two thousand years ago, Nhagruul the Foul, a sorcerer who revealed....	violence	imdb
tt0033045	The Shop Around the Corner	Matuschek's, a gift store in Budapest, is the workplace...	romantic	imdb

Table 3: Dataset details: Custom Chatbot Corpus (Smoking Cessation domain).

ID	Title	Free_text	Tags
01	Smoking	I had alcohol	Alcohol
02	Vaping	Can you please remove me from this program	exit
03	Smoking	I am in a bad mood	Badmood, mood

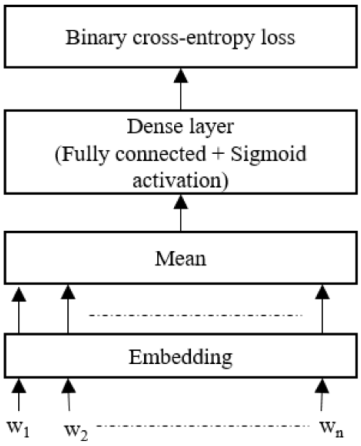

Figure 2: CBOW Model architecture

Model details

Word2Vec [20] framework consists of a group of related models that give word vectors. Word2Vec is one of the most popular technique proposed by Google to learn word embeddings with neural networks. The embeddings can be obtained using two algorithms: Skip-Gram [21] and Continuous Bag of Words (CBOW) [17]. For learning, the neural network in both the algorithms uses back-propagation. Both of these algorithms differ in whether the neural network tries to predict a focus word from the surrounding context words as in CBOW or the opposite.

In this work, we utilize CBOW as an input for classification with shallow neural network. While CBOW like conventional bag of words lose the information about the order of words. However, word embeddings for each word created with word as the target gives CBOW an added advantage. CBOW can also pick the transitions from previously unseen words. Another advantage is that irrespective of the length of user utterance or vocabulary, the input is a fixed dimensional vector for the classification method. CBOW first finds either the sum or mean vector value of the surrounding words, and then projects that value to a position in the vector space. The summation of the surrounding word vectors results in a path in the vector space. The resultant vector captures the numerical representation of user utterance's meaning.

$$CBOW_{mean}(w_1, w_2, ..., w_n) = \frac{1}{n}\sum_{i=1}^{n} v(w_i) \quad (1)$$

If more words with the same meaning are added, the classification might be harder due to the spread of a keyword's cluster of representations. When the average is used, the order of the nearby words does not matter, and the length of the path is normalized with respect to the word count of the user input. Fig. 2 represents CBOW architecture with an embedding layer that is applied to all the words in user input (eq. 1), where w_i is a one-hot encoding vector of the vocabulary. The fixed-size vector obtained by averaging the embedded words is fed to the neural network. The model is trained on CBOW architecture with key parameters that include position weights, 300 embedding dimensions, and a context window length of 5 words.

The neural network model consists of 3 layers: First input dense layer with 128 neurons, second hidden layer with 64 neurons, both with rectified linear unit (ReLU) non-linear activations and the third Sigmoid output layer consists of the number of neurons equal to the number of keywords in the dataset. The resultant fixed dimensional vector from CBOW is fed to the above-mentioned shallow neural network to transform the vector into the probabilities. A fixed threshold is chosen to determine whether to assign a particular keyword. Thus, CBOW with neural network predicts the most probable keyword by considering the complete user input.

Evaluation Results and Discussion

Case-study Chatbot Model

This section describes a closed-domain chatbot model considered for keyword detection evaluation in this paper. The chatbot model is specifically tailored to the healthcare intervention domain. The knowledge base consists of 6 months long protocol (manuals developed by healthcare professionals), structured data as well as text corpus, and other relevant sources of information on smoking cessation. Fig. 3 shows an example conversation with the rule-based smoking cessation chatbot model. Typically, closed-domain chatbots like the one in this work are highly effective at handling specific tasks with their domain of expertise. Thus, also providing quick and accurate responses to frequently encountered and common user queries. These characteristics make the closed-domain chatbot require least human intervention.

However, closed-domain chatbots could be limited in their ability to handle out of vocabulary queries and struggle with complex, and ambiguous user inputs. This necessitates the integration of NLP techniques to improve the accuracy of understanding free-text from users. Fig. 4 illustrates the conversation of smoking

cessation chatbot embedded with NLP-based keyword detection with an example of extracting the actual numeric from the text form of a number. This paper presents the performance evaluation of the keyword detection (methodology discussed in section III) with limited keywords.

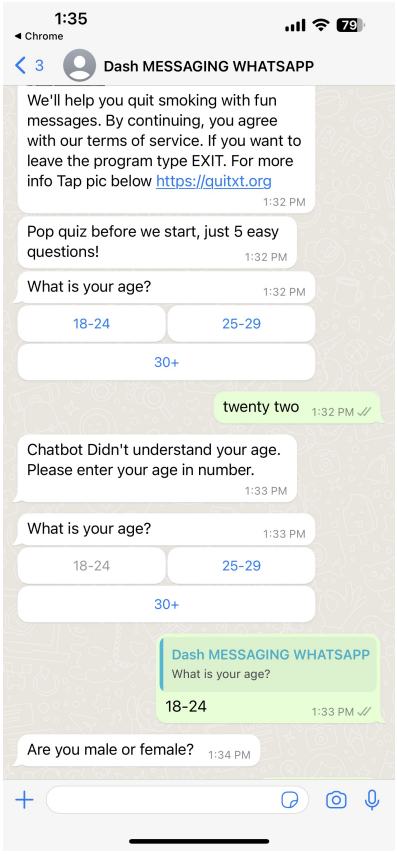

Figure 3: Case study chatbot dialog with purely rule-based predefined keywords

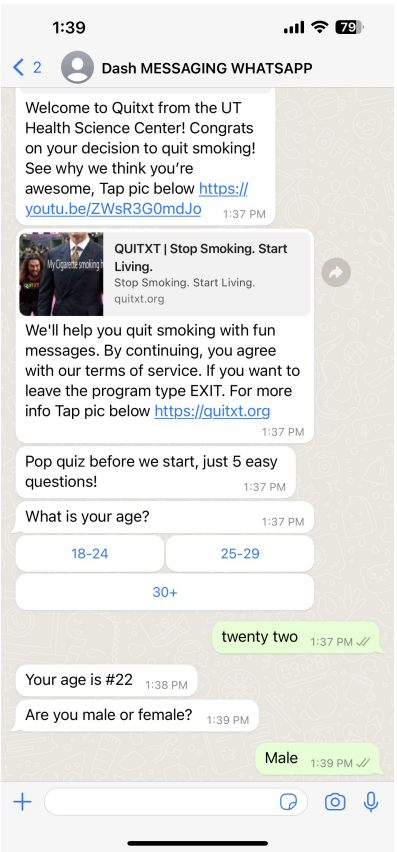

Figure 4: Case study rule-based chatbot dialog with custom NLP in keyword detection

Results

We validated the results of the NLP model with Movie Tag and custom chatbot corpus datasets with a few measures that are common for evaluating multi-label text classifiers. Since our custom dataset is unbalanced in terms of examples per keyword, we obtain the micro-averaged test scores via the following metrics:

- Precision: A ratio of correctly predicted positive observations to the total positive observations prediction. Hence, it is the number of true positives over the sum of the number of true positives and the number of false positives.
- Recall, also known as sensitivity, is the ratio of correctly predicted positive observations to all the actual positive observations. Hence, it is the number of true positives over the sum of the number of true positives and the number of false negatives.
- F1-score is calculated from the weighted harmonic mean of precision and recall. F1-score takes into account both the false positives and false negatives. It proves useful in cases where we have an uneven distribution of utterances per keyword.

A total of 3 sets of keyword counts on Movie Tag and custom chatbot datasets. The proposed CBOW model is fine-tuned on the two datasets. Tables 4 and 5 show the classification task performance results for Movie tags and custom chatbot corpus datasets. The performance slightly improves when reducing the count based on keyword reduction approach. Model converges in a lesser number of epochs which indicates the usefulness of closed-domain chatbots with limited keywords. F1 score improves with reduced keyword count on Movie Tag dataset while exhibiting comparable performance with custom chatbot dataset. The keyword detection performance is additionally attributed to the quality of the keywords.

While most of the generic-domain NLP services such as Amazon Lex [22], Google Dialogflow [23], Microsoft LUIS [24], perform better since they are fed with hundreds if not thousands of user responses everyday. The closed-domain chatbots (as discussed in this study) do not need to be trained on such big datasets, instead require the specific task-based data. The custom dataset is task-specific and are curated based on the inputs, domain expert knowledge, previous user response logs, and protocol developed by healthcare professionals.

Table 4: Keyword extraction performance with Movie Tag dataset.

Keyword count	Precision	Recall	F1-Score
100	0.87	0.86	0.87
50	0.89	0.88	0.88
40	0.92	0.9	0.91

Table 5: Keyword extraction performance with Chatbot Corpus dataset.

Keyword count	Precision	Recall	F1-Score
28	0.8	0.78	0.79
14	0.82	0.8	0.81
7	0.83	0.79	0.81

Conclusion

Closed-domain chatbots are typically topic-centric, despite longer and deep conversational dialog flow. Such chatbots can be trained on domain-specific data. Thus, keeping frequently used keywords for the task could alleviate the need of utilizing huge datasets (typical of NLP models for improved performance) and suitable for resource-constrained setting. The rule of thumb is that the task-based bots mostly follow pre-defined similar dialog turns throughout the complete duration of a conversation. This work presents a preliminary evaluation results with the keyword reduction approach on closed-domain chatbot datasets. The experimental results provide direction in choosing subsets of original corpus for a specific domain without performance degradation. In the future work, we intend to perform an in-depth evaluation of the keyword reduction approach on wider domain datasets.

References

[1] "Chatbot market," https://www.alliedmarketresearch.com/chatbot-market, accessed: 2021-09-30.

[2] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 1371–1374.

[3] "A.l.i.c.e bot," https://en.wikipedia.org/wiki/Artificial_Linguistic_Internet_Computer_Entity, accessed: 2021-04-30.

[4] "Cleverbot," https://en.wikipedia.org/wiki/Cleverbot, accessed: 2021-04-30.

[5] "Jabberwacky," http://www.jabberwacky.com/default.aspx, accessed: 2021-04-30.

[6] "Mitsuku," https://steemit.com/steemhunt/@juecoree/mitsuku-the-world-s-best-conversational-chatbot, accessed: 2021-04-30.

[7] H. Io and C. Lee, "Chatbots and conversational agents: A bibliometric analysis," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE, 2017, pp. 215–219.

[8] R. Khan and A. Das, "Build better chatbots," *A complete guide to getting started with chatbots*, 2018.

[9] A. G. Ramirez, P. Chalela, D. Akopian, E. Munoz, K. J. Gallion, C. Despres, J. Morales, R. Escobar, and A. L. McAlister, "Text and mobile media smoking cessation service for young adults in south texas: operation and cost-effectiveness estimation," *Health promotion practice*, vol. 18, no. 4, pp. 581–585, 2017.

[10] P. Jackson and I. Moulinier, *Natural language processing for online applications: Text retrieval, extraction and categorization*. John Benjamins Publishing, 2007, vol. 5.

[11] H.-Y. Shum, X.-d. He, and D. Li, "From eliza to xiaoice: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 10–26, 2018.

[12] A. Kulshreshtha, D. D. F. Adiwardana, D. R. So, G. Nemade, J. Hall, N. Fiedel, Q. V. Le, R. Thoppilan, T. Luong, Y. Lu *et al.*, "Towards a human-like open-domain chatbot," 2020.

[13] J. L. Z. Montenegro, C. A. da Costa, and R. da Rosa Righi, "Survey of conversational agents in health," *Expert Systems with Applications*, vol. 129, pp. 56–67, 2019.

[14] S. Larson, A. Mahendran, J. J. Peper, C. Clarke, A. Lee, P. Hill, J. K. Kummerfeld, K. Leach, M. A. Laurenzano, L. Tang *et al.*, "An evaluation dataset for intent classification and out-of-scope prediction," *arXiv preprint arXiv:1909.02027*, 2019.

[15] "The state of chatbots in 2019," https://medium.com/hackernoon/the-state-of-chatbots-in-2019-d97f85f2294b, accessed: 2021-05-30.

[16] R. Pieraccini and J. M. Huerta, "Where do we go from here? research and commercial spoken dialogue systems," *Recent trends in discourse and dialogue*, pp. 1–24, 2008.

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

[18] K. Lu, X. Cai, I. Ajiferuke, and D. Wolfram, "Vocabulary size and its effect on topic representation," *Information Processing & Management*, vol. 53, no. 3, pp. 653–665, 2017.

[19] "Movie plot synopses data," https://www.kaggle.com/cryptexcode/mpst-movie-plot-synopses-with-tags, accessed: 2022-05-30.

[20] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[21] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks, "A closer look at skip-gram modelling." in *LREC*, vol. 6, 2006, pp. 1222–1225.

[22] "Amazon lex," https://aws.amazon.com/lex/, accessed: 2022-09-30.

[23] "Google dialogflow," https://developers.google.com/learn/pathways/chatbots-dialogflow, accessed: 2022-09-30.

[24] "Microsoft luis," https://www.luis.ai/, accessed: 2022-09-30.