Building Kurdish Chatbot Using Free Open Source Platforms

Kanaan M. Kaka-Khan

Department of Computer Science, University of Human Development, Iraq

ABSTRACT

Chatbot is a program that utilizes natural language understanding and processing technology to have a human-like conversation. Nowadays chatbots are capable to interact with users in world's majority languages. Unfortunately, bots that interact with Kurdish users are rare. This paper is an attempt to bridge the gap between chatbots and Kurdish users. This paper tries to implement a free open source platform (pandorabots) to build a Kurdish chatbot. I present a number of challenges for Kurdish chatbot at the last section of this work.

Index Terms: Artificial Intelligence, Artificial Intelligence Markup Language, Chatbot, Pandorabots

1. INTRODUCTION

A. Chatbot

A chatbot is a service, powered by rules and sometimes artificial intelligence that you interact with via a chat interface [1,2]. They range from simple systems that extract a response from databases when they match certain keywords to more sophisticated ones that use natural language processing techniques [3].

B. Needs for Chatbot

And an extraordinary focus was devoted to chatbots within the tech community in recent years [4]. There is no doubt that majority of business are going to be online; if we want to make a business online we have to locate where the people are? That place now is the zone of messenger applications as mentioned by Peter Rojas "People are now spending more time in messaging apps than in social media and that is a huge turning point. Messaging apps are the platforms of the future and bots will be how their users access all sorts of services" [5]. Any user's interaction with an app or web page can utilize a Chatbot to increase the user's experience [6].

Fig. 1 shows the size of the top 4 messaging apps and social networks; big 4 messaging apps are Whatsapp, Messenger, WeChat, Viber, big 4 social networks are Facebook, Instagram, Twitter, and LinkedIn [7].

C. Applications of Chatbot

The very basic use at the early days of chatbot was almost restricted to conversations. The first chatbot in history was Eliza, a program which represents a psychologist [8]. By the time the bot provides a wide range to many important applications, some of the most important applications of chatbots are listed below:

1. Customer service
2. Mobile personal assistants
3. Advertisements
4. Games and entertainment applications
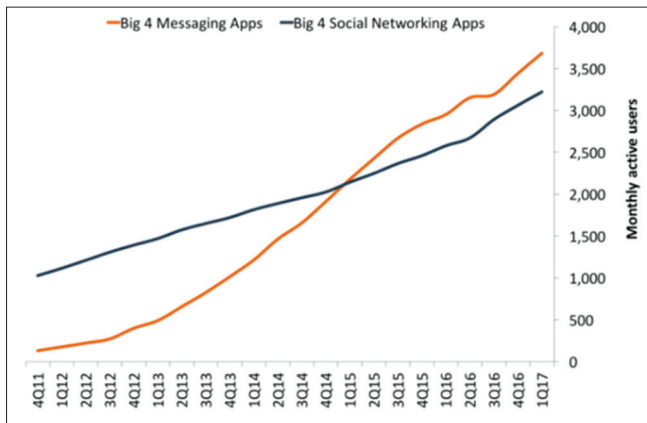5. Talking toys
6. Call centers.

Fig. 1. Users for top 4 messaging apps and social networks in million [7]

The crucial aim of this work is to build a bot that is capable of working as a guide who is sitting on the UHD website and giving information about the University of Human Development to any user whenever asked.

2. CHATBOT HISTORY

The concept of natural language processing generally and chatbots specifically can be originated to Alan Turing question "Can machines think?" who asked in 1950 [9]. Alan's question (which is called Turing Test now) is nothing just asking questions to human and machine subjects, to identify the human. We say the machine can think if the human and machine responses are indistinguishable. In 1966, Eliza (the first chatbot) was created by Joseph Weizenbaum at MIT. For generating proper responses, Eliza uses a set of pre-programmed rules to identify keywords and pattern match those keywords from an input sentence [8]. In 1995, a new more complex bot (A.L.I.C.E) created by Richard Wallace. ALICE makes use of artificial intelligence markup language (AIML) to represents conversations as sets of patterns (inputs) and templates (outputs). ALICE got Loebner prize (yearly chatbot competition) thrice and award the most intelligent chatbot [10]. Advances in natural language processing and machine learning played important roles in improving chatbot technology; modern chatbots include Microsoft's Cortana, Amazon's Echo and Alexa, and Apple's [11].

3. RELATED WORKS AND METHODOLOGY

As in many natural language processing applications, there are many approaches to developing chatbot: Using a set of predefined rules [12], semi automatically learning conversational pattern from data [13], and full automatic chatbot (under researching). Each approach has its own merits and demerits, through manual approach more control over the language and the chatbot can be achieved, but it needs more effort to maintain a huge set of rules. The second approach which also is called corpus-based is challenged by the need to construct coherent personas using data created by different people [Botta]. Due to lack of Kurdish corpus (at least it is not available for me even if it exists), I chose manually written rules by making use of AIML, a popular programming language to represents conversations as a set of patterns (inputs) and templates (outputs).

As in other NLP applications, in the area of Kurdish chatbot, unfortunately, we find related works rarely. With the best of my knowledge this is the first Kurdish chatbot which is created academically, so sometimes I obliged to relate my work with Arabic or Persian languages. Most notably, in 2016, Dana and Habash developed Botta, the first Arabic dialect Chatbot, Botta explore the challenges of creating a conversational agent that aims to stimulate friendly conversations using the Egyptian Arabic dialect [3].

Playground and programming language are the two basic requirements for creating chatbots. Playground can be defined as a sandbox or an integrated development environment for the programming language [1]. In this work, I chose pandorabots as a playground (creating, deploying, talking with the bot) and AIML (for Making conversation) as a programming language for creating Kurdish chatbot, ALICE, an award-winning free chatbot was created using AIML [12].

After login into pandorabots playground with Facebook account, the work will be shown in the following steps:
- Step 1: I gave "kuri zanko" as the bot name.
- Step 2: In the bot editor space, I created a file named "UHD" which is AIML file to involve all the patterns (inputs) and templates (outputs).
- Step 3: I started writing an expected user input in <pattern></pattern> tag and the bot answer in <template></template> tag, both pattern and template are enclosed in a <category></category>, a category is the basic unit of knowledge in AIML [1].
- Step 4: After writing each category, I train (test) the bot to know whether it gives the correct answer.
- Step 5: After writing all the categories, the bot will be published in the pandorabots clubhouse (a public place where users can talk to the bots).

4. RESULT AND DISCUSSION

For the simple and direct user input the bot can give the answer easily, for example:

User: سڵاو

Bot: سڵاو له بهڕێزتان،خۆتان بناسێنن

A. Pattern Matching
To form a user input matching, the bot searches through its AIML file (categories). It may happen, a user input does not match any of the pattern defined in our bot, so a default answer should be provided which is called ultimate default category:

<pattern>*</pattern>

<template>

ببوره بهڕێزم، وهڵامی پرسیارهکهتم لانیه

</template>

The star (*) determines that a user input does not match any of the bot patterns, relying on one default answer is extremely tedious for the clients. This obliges us to think about random responses to provide different responses for the same user input.

<random>

ببوره بهڕێزم، وهڵامی پرسیارهکهتم لانیه

بهڕێزم پرسیارهکهت بهجۆرێکی تر بکهرهوه

بهڕێزم پرسیارهکهت رون نیه

ببوره له پرسیارهکهت نهگهیشتم

</random>

These random responses make sense that the user is chatting with a human, not a bot.

B. Wildcards
Wildcards are used to capture many inputs using only a single category [1]. Through wildcards bots can be more intelligence. There are many wildcards but (* and ^) are the most two ones which are used in this work:

<pattern> ناوم *</pattern

In this example, the star(*) stands for any name that is given by the user.

<pattern> زانکۆی گهشهپێدان *</pattern>

In the second example, the star stands for any words or sentences which appear after the name "زانکۆی گهشهپێدان".

<pattern> ^ کۆمپیوتهر ^</pattern>

The (^) wildcard lets the bot to capture any input containing the word "کۆمپیوتهر" and gives a proper answer.

Wildcards should be used carefully because their priority is different, Fig. 2 shows wildcard and exact matching priorities.

A category with # wild card will be matched first and * wildcard will be matched last, for example: When a user even types "سڵاو له ئێوه"the response will be taken from "سڵاو #" pattern not "سڵاو له ئێوه" pattern.

C. Variables
Bot intelligence can also be achieved through variables. Variables can be used to store information about your bot and the users; this gives the user a sense that he/she is chatting with a human being. Fig. 3 shows a short conversation between my bot and a user.

D. Recursion
Recursion means writing a template that is calling another category, and this leads to minimizing the number of categories in our bot AIML file.

<pattern>های</pattern>

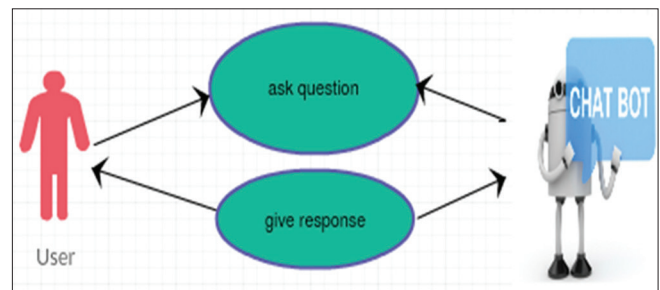<template><srai>سڵاو</srai></template>

Fig. 2. Chatbot simple flow diagram

سڵاو * > سڵاو ^ > سڵاو له ئێوه > سڵاو _ > سڵاو #

Fig. 3. Wildcards priority

Through using recursion, no need to rewrite a new category to input "های", we just refer to the template "سلاو" using <srai> tag, and the bot answers the user exactly as he/she said "سلاو" to the bot.

E. Context

To make our bot capable of doing human-like conversation, it should remember the things that have been previously said. My bot is capable of remembering the last sentence it said. (Fig. 4-6) shows different conversations regarding context.

F. Challenges

- Challenge 1: The first and greatest challenge for Kurdish Chatbot is the lack of platform designed specifically to Kurdish Language, Kurdish structure extremely differs from English or any other languages, Kurdish word order is SOV [subject+ object+ verb] [14]. The reason behind

the slow progress in Arabic NLP is the complexity of the Arabic language [3], same to Kurdish. Hence, it is very tough to have a very intelligent Kurdish bot using free open source platforms.

- Challenge 2: Dialectal Variation, Kurdish language has many different dialects; the gap among dialects sometimes reaches a level that speakers of a dialect do not understand another dialect, and it means that it is quite tough to build a bot capable of chatting with all different Kurdish dialects.

- Challenge 3: Normalization is one of the important processes in developing bots, normalization includes sentence splitting, correcting spelling errors, person, and gender substitution.

wanna -> want to

isn't -> is not

How R U -> How Are You

With you -> with me

The user may be bad in spelling, he/she may type "how r u" instead of "how are you". These changes (normalization

Fig. 4. A sample conversation between a user and the bot

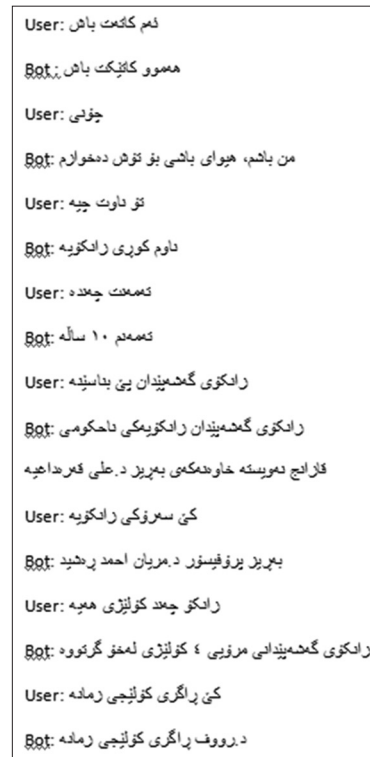**Fig. 5.** A sample conversation regarding context

Fig. 6. Detailed conversation between a user and the bot

and substitution) can be done easily in English and make the bot to interact with the user as a human not a bot, while it's a bit difficult to perform the same for Kurdish because the bot components (AIML files, Set files, and Map Files) are already exist for English language while not for Kurdish, it requires vast effort from both computer science and linguistic people to maintain such files.

- Challenge 4: In spite of majority of platforms claiming for language agnosticism, practically we face issues for Kurdish due to its own structure. For example, when a name is given, as "Alan" to the bot and later on he asks the bot about his name it says "your name is Alan." While the same name is given in Kurdish language"ئالان" to the bot and I ask the bot for his name, it should tell "تۆ ناوت ئالانه" a suffix will be seen "ه" with the name "ئالان", this seems to be an easy task but really needs a hard work to do.

5. CONCLUSION AND FUTURE WORK

Chatbots are online human-computer dialog system[s] with natural language [15]. I have presented the first Kurdish chatbot and described some of the challenges for Kurdish chatbot. Building chatbot from scratch is extremely tough, time consuming, costly. This reason led me to go for free open source platform (pandorabots). This work aims to be a basic structure for Kurdish dialect, providing future Kurdish bot masters with a base chatbot which contains basic files, general knowledge.

6. BIOGRAPHY

Kanaan M. Kaka-Khan is an associate professor in the Computer Science Department at Human Development University, Sulaimaniya, Iraq. Born in Iraq 1982. Kanaan M. Khan had his bachelor degree in Computer Science from Sulaimaniya University, and Master Degree in IT from BAM university, India. His research interest area includes Natural Language Processing, Machine Translation, Chatbot, and Information Security.

REFERENCES

[1] "How to Build a Bot using the Playground UI". Available: https://www.playground.pandorabots.com/en/tutorial. [Last Accessed on 2017 Aug 25].

[2] "The Complete Beginner's Guide to Chatbots." Matt Schlicht, Founder of Chatbots Magazine, Apr. 20, 2016. Available: https://www.chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca. [Last Accessed on 2017 Aug 25].

[3] "Botta: An Arabic Dialect Chatbot." Dana Abu Ali and Nizar Habash, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, pp. 208-212, Dec. 11, 17, 2016.

[4] "Best uses of Chatbots in the UK." Charlotte Jee. Available: http://www.techworld.com/picture-gallery/apps-wearables/9-best-uses-of-chatbots-in-business-in-uk-3641500. Jun. 08, 2017.

[5] "Chatbot Survey 2017." Ayush Jain, Co-founder and CEO at Mindbowser. Available: https://www.slideshare.net/Mobileappszen/chatbots-survey-2017-chatbot-market-research-report. [Feb. 08, 2017.

[6] "Chatbot Applications and Considerations." Josef Ondrejcka. Available: http://ramseysolutions.com/chatbot-applications-and-considerations. [Sep. 19, 2016].

[7] "Messaging Apps are Now Bigger than Social Networks." BI Intelligence. Available: http://www.businessinsider.com/the-messaging-app-report-2015-11. [Sep. 20, 2016].

[8] J. Weizenbaum. "ELIZA-a computer program for the study of natural language communication between man and machine." *Communications of the ACM,* vol. 9, no. 1, pp. 36-45, 1966.

[9] A. M. Turing. "Computing machinery and intelligence." *Mind*, vol. 59, no. 236, pp. 433-460, 1950.

[10] R. S. Wallace. "The Anatomy of A.L.I.C.E." Available: http://www.alicebot.org/anatomy.html. [Last Accessed on 2017 Aug 25].

[11] M. Weinberger. Why Amazon's Echo is Totally dominating-and what Google, Microsoft, and Apple have to do to Catch Up. Available: http://www.businessinsider.com/amazon-echo-google-home-microsoft-cortana-apple-siri-2017-1. [Jan. 14, 2017].

[12] R. Wallace. *The Elements of AIML Style,* San Francisco: Alice AI Foundation, 2003.

[13] B. A. Shawar and E. Atwell. "Using dialogue corpora to train a chatbot." In Proceedings of the Corpus Linguistics 2003 Conference, pp. 681-690, 2003.

[14] "Evaluation of in Kurdish machine translation system." Kanaan and Fatima, Proceedings of UHD 2017, the 4th International Scientific Conference, Sulaimanya, Iraq, pp. 862-868, Jun. 2017.

[15] J. Cahn. "CHATBOT: Architecture, design, and development." University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, Apr. 26, 2017.