Botta: An Arabic Dialect Chatbot

Dana Abu Ali and Nizar Habash
Computational Approaches to Modeling Language Lab
New York University Abu Dhabi, UAE
`{daa389,nizar.habash}@nyu.edu`

Abstract

This paper presents BOTTA, the first Arabic dialect chatbot. We explore the challenges of creating a conversational agent that aims to simulate friendly conversations using the Egyptian Arabic dialect. We present a number of solutions and describe the different components of the BOTTA chatbot. The BOTTA database files are publicly available for researchers working on Arabic chatbot technologies. The BOTTA chatbot is also publicly available for any users who want to chat with it online.

1 Introduction

Chatbots are conversational agents that are programmed to communicate with users through an intelligent conversation using a natural language. They range from simple systems that extract responses from databases when they match certain keywords to more sophisticated ones that use natural language processing (NLP) techniques. These conversational programs are commonly used for a variety of purposes from customer service and information acquisition to entertainment.

While chatbots remain English-dominated, the technology managed to spread successfully to other languages. The Arabic language is one of the under-represented languages in many NLP technologies, and chatbots are not an exception. The reason behind the slow progress in Arabic NLP is the complexity of the Arabic language, which comes with its set of challenges such as very rich morphology, high degree of ambiguity, common orthographic variations, and numerous dialects.

In this paper we present, BOTTA,[1] an Arabic dialect chatbot. BOTTA is a conversational companion that uses the commonly understood Egyptian Arabic (Cairene) dialect to simulate friendly conversation with users. To our knowledge, BOTTA is the first chatbot in an Arabic dialect. The BOTTA database files are publicly available for researchers working on Arabic chatbot technologies.[2] The BOTTA chatbot is also publicly available for any users who want to chat with it online.[3]

The rest of this paper is organized as follows. Next, we present related work in Section 2. We discuss the challenges for developing Arabic chatbots in Section 3. We then describe our approach and design decisions in some detail in Section 4 including presenting a conversation example and a preliminary user evaluation.

2 Related Work

As in many other NLP areas, there are two types of approaches to developing chatbots: using manually written rules (Wallace, 2003) or automatically learning conversational patterns from data (Shawar and Atwell, 2003; Sordoni et al., 2015; Li et al., 2016). Both approaches have some advantages and disadvantages. While manual rules allow more control over the language and persona of the chatbot, they are

---

[1] The name BOTTA (pronounced like but-ta) evokes the English word Bot as well as the Arabic friendly female nickname Batta (بطّة), which can be translated as ‘Ducky’.

[2] To obtain the BOTTA database files, go to `http://camel.abudhabi.nyu.edu/resources/`.

[3] To chat with BOTTA online, go to `https://playground.pandorabots.com/en/clubhouse/` and search for ‘Botta’. The current version (1.0) is under BotID `dana33/botta`.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 208–212, Osaka, Japan, December 11-17 2016.

tedious to create and can be at times unnatural. Corpus-based techniques are challenged by the need to construct coherent personas using data created by different people. For BOTTA, we chose not to start with the corpus-based approach because we wanted to model certain aspects of the complexity of the Arabic language and address its challenges in a more controlled setting. We plan to make use of existing conversational corpora, such as Call Home Egyptian (Gadalla et al., 1997), in the future. But we are aware of many challenges: Call Home consists of recorded conversations between acquaintances, and BOTTA is supposed to talk to strangers; and not to mention that the speakers in Call Home vary in age and gender, while BOTTA is supposed to be a young woman. Another option is to get data from forums and Twitter, but the problem with that is that BOTTA will be borrowing different people's perspectives and personal opinions, which holds the risk of making her sound incoherent.

In the area of Arabic chatbots, little work has been done. Most notably, Shawar and Atwell (2004) developed a chatbot for question answering over Quranic text. Shawar (2011) described a question answering system focusing on the medical domain. Both of these efforts are in Standard Arabic (the first technically in Classical Arabic). AlHagbani and Khan (2016) discussed a number of challenges facing the development of an Arabic chatbot. They describe the challenges in detail but briefly describe the development of a simple chatbot.

In developing BOTTA, we make use of AIML (Artificial Intelligence Markup Language), a popular language used to represent dialogues as sets of patterns (inputs) and templates (outputs). ALICE, an award-winning free chatbot was created using AIML (Wallace, 2003). There are thousands of adaptations of ALICE made by botmasters who use her software as the base of their chatbots. One variation of ALICE is Rosie, a chatbot that was optimized for use on the Pandorabots online platform.[4] BOTTA aims to become the Rosie of Arabic dialects, providing future Arab botmasters with a base chatbot that contains basic greetings, general knowledge sets, and other useful features.

In the next section, we present a summary of the challenges we encountered developing BOTTA.

3 Arabic Natural Language Processing Challenges

The following are some of the main Arabic NLP challenges, with a focus on chatbots.

Dialectal Variation Arabic consists of a number of variants that are quite different from each other: Modern Standard Arabic (MSA), the official written and read language, and a number of dialects, the spoken forms of language (Habash, 2010). While MSA has official standard orthography and a relatively large number of resources, the various Arabic dialects have no standards and only a handful of resources. Dialects vary from MSA and each other in terms of phonology, morphology and lexicon. Dialects are not recognized as languages and not taught in schools in the Arab World. However, dialectal Arabic is commonly used in online chatting. This is why we find it more appropriate to focus on dialectal Arabic in the context of a chatbot. While BOTTA speaks in the Cairene Egyptian Arabic dialect, she recognizes common words and greetings in a number of other dialects.

Orthographic Ambiguity and Inconsistency Arabic orthography represents short vowels and consonantal doubling using optional diacritical marks, which most commonly are not included in the text. This results in a high rate of ambiguity. Furthermore, Arabic writers make very common mistakes in spelling a number of problematic letters such as Alif-Hamza forms and Ta-Marbuta (Zaghouani et al., 2014). The issue of orthography is exacerbated for Arabic dialects where no standard orthographies exist (Habash et al., 2012; Eskander et al., 2013).

Morphological Richness Arabic words are inflected for a large number of features such as gender, number, person, voices, aspect, etc., as well as accepting a number of attached clitics. In the context of a chatbot system this proves very challenging. Verbs, adjectives, and pronouns are all gender specific, which requires the chatbot to have two different systems of responses – one for male users and another for female users.

---

[4] `http://www.pandorabots.com/`.

Idiomatic Dialogue Expressions As with any other language, Arabic has its own set of unique idiomatic dialogue expressions. One common class of such expressions is the modified echo greeting responses, e.g., while the English greeting ‘Good Morning’ gets an echo response of ‘Good Morning’, the equivalent Arabic greeting صباح الخير *SbAH Alxyr*[5] ‘lit. Morning of Goodness’ gets a modified echo response of صباح النور *SbAH Alnwr* ‘lit. Morning of Light’.

Because of all these challenges, an Arabic-speaking chatbot requires its unique databases, as opposed to a machine translation wrapper around an existing English-speaking chatbot.

Next, we discuss Botta’s design and components.

4 Botta

Botta’s persona is that of a friendly female chatbot, who aims to simulate conversation and connect with as many Arab users as she can. She is the first chatbot that converses in an Arabic dialect, which supports her purpose of entertaining users who are accustomed to chatting in the dialect. We created Botta using AIML and launched it on the Pandorabots platform. Botta’s knowledge base is made up of AIML files that store the categories containing its responses to the user inputs, set files of themed words and phrases, and map files that pair up related words and phrases.

4.1 AIML Files

The main file here is the greetings file, which divides the basic greetings in a number of Arabic dialects into categories. The templates, or responses, in these categories also contain questions that allow Botta to learn basic information about the user, such as age, gender, and nationality. Botta retrieves that information when needed, such as when formulating gender-inflected responses. Another AIML file stores Botta’s bio, explaining her background to users and asking them questions about themselves in return to maintain conversations. Other AIML files are linked to map and set files, which will be discussed next. The categories in the AIML files have certain patterns that would initiate the extraction of information from the other files.

4.2 Set Files

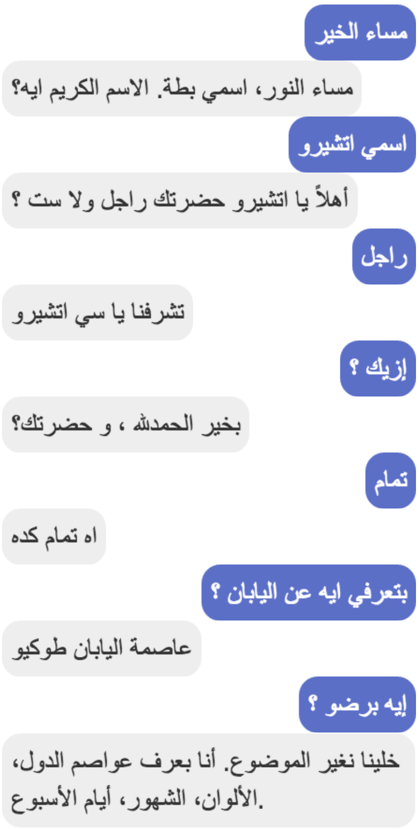
Sets in AIML are simple lists that are used to store words and phrases that fall under one theme. Botta has been equipped with certain lists that provide her with some general knowledge, which she can use to continue the conversation and entertain users. Some of these sets have been directly translated from Rosie’s sets, such as the countries’ set. Other sets have been modified according to what is used in Arabic dialects. For example, months in Arabic have different names depending on the regional dialect; therefore, Botta has separate sets of months based on each dialect it recognizes. Botta also has unique sets, storing dialectal bad language that offend her, words that indicate the dialect of the users, and other regional knowledge.

4.3 Map Files

Maps in AIML are lists of key-value pairs that are used to relate words to certain words and phrases. Keys in maps should be stored in sets for the mapping to work. Sets and maps have to be called within the AIML files; otherwise, they remain idle in the chatbot memory. Botta has a `nation2capital` map, a translated version from the same Rosie file. Botta also has the `word2proverb map`, which allows her to generate a funny Arabic dialect proverb when she matches on a word but does not understand the phrase. One of the most important maps Botta has is the `name2gender` map, which is used to determine the user’s gender, and thus the gender-inflected responses. AIML files are used to activate sets and maps. One example of this is the `guessingthegender.aiml` file, which would be used when a user provides Botta with a name. Botta searches for the name in the `names.set`; if the name is found, she searches for the corresponding gender in the `name2gender.map` file. After confirming

---

[5] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007) .



مساء الخير	U: Good evening (lit. evening of goodness)
مساء النور، اسمي بطة. الاسم الكريم ايه؟	B: Good evening (lit: evening of light), your good name?
اسمي اتشيرو	U: My name is Ichiro
أهلاً يا اتشيرو حضرتك راجل ولا ست ؟	B: Hello Ichiro. Are you a man or a woman?
راجل	U: Man
تشرفنا يا سي اتشيرو	B: Nice meeting you Mr. Ichiro
إزيك ؟	U: How are you?
بخير الحمدلله ، و حضرتك؟	B: Good, *Alhamdulillah* (thanks to God), and you?
تمام	U: Alright
اه تمام كده	B: That's good
بتعرفي ايه عن اليابان ؟	U: What do you know about Japan?
عاصمة اليابان طوكيو	B: Japan's capital is Tokyo.
إيه برضو ؟	U: What else?
خلينا نغير الموضوع. أنا بعرف عواصم الدول، الألوان، الشهور، أيام الأسبوع.	B: Let's change the subject. I know capitals, colors, months, and weekdays.

Figure 1: A sample conversation between a user (U) and BOTTA (B).

the guessed gender with the user, BOTTA sets the gender variable and starts using the correct gender-inflected responses during the conversation. If the name is not found in the `names.set` file, BOTTA asks for that information from the user instead – see the example conversation in Figure 1, where BOTTA does not recognize the user's name.

4.4 Orthography Handling

In BOTTA, we use orthographic normalization to overcome the inconsistent spelling variations of certain characters. Some of the changes we made were borrowed from the Conventional Orthography of Dialectal Arabic (Habash et al., 2012), which is an internally consistent and coherent convention for writing dialectal Arabic. While BOTTA does not use normalized text in its responses, changing user inputs to that form enhances the matching liklihood. Rosie uses this technique for reductions and contractions. BOTTA performs the following orthographic transformations:

- The word-final Alif-Maqsura letter ى *ý* is often used in Egypt to write word-final Ya ي *y*, and vice versa (Eskander et al., 2013). In BOTTA, we change every Alif-Maqsura to Ya.
- We change every Ta-Marbuta ة *ħ* to Ha ه *h*, since these two letters are often confused in word-final positions.
- The misspelling of the Alif-Hamza forms (أ Â, إ Ǎ and ا A) is the most common spelling mistake in Arabic, with a 38.5% frequency rate (Eskander et al., 2013). We normalize all Alif-Hamza forms (أ Â, إ Ǎ) to bare Alif (ا A).

By performing the above-mentioned transformations, BOTTA's pattern matching will be able to overcome 85.1% of the spelling mistakes found in spontaneous Arabic typing (Eskander et al., 2013).

4.5 Preliminary User Evaluation

We asked three native Arabic speakers to chat with BOTTA and evaluate the naturalness of the conversation. Two of them are native Egyptian Arabic speakers, and one is a Levantine Arabic speaker. They all agreed that they found her entertaining and wanted the conversation to last longer. They commented that her Egyptian Arabic sounds authentic. Not being informed of BOTTA's purpose beforehand, they all guessed that she was created to carry out a conversation and not perform tasks. They pointed out that she gets repetitive sometimes, and makes out-of-context statements. Their suggestions include having her talk about herself more, asking the user more questions, and leading the conversation by introducing new topics.

5 Conclusions and Future Work

We have presented BOTTA, the first Arabic dialect chatbot and described the challenges and some solutions to building chatbots in Arabic. The BOTTA files are publicly available for researchers working on Arabic chatbot technologies. In the future, we plan to enhance BOTTA's pattern matching using corpus-based machine learning techniques. Further development will also include exploiting existing tools for Egyptian Arabic processing (Pasha et al., 2014) to perform morphological analysis on the input and to experiment with lemma-based pattern matching.

References

Eman Saad AlHagbani and Muhammad Badruddin Khan. 2016. Challenges facing the development of the Arabic chatbot. In *First International Workshop on Pattern Recognition*, pages 100110Y–100110Y. International Society for Optics and Photonics.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic Transcripts. In *Linguistic Data Consortium, Philadelphia*.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash, Mona Diab, and Owen Rabmow. 2012. Conventional Orthography for Dialectal Arabic. In *Proceedings of LREC*, Istanbul, Turkey.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Arfath Pasha, Mohamed Al-Badrashiny, Ahmed El Kholy, Ramy Eskander, Mona Diab, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *In Proceedings of LREC*, Reykjavik, Iceland.

Bayan Abu Shawar and Eric Atwell. 2003. Using dialogue corpora to train a chatbot. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 681–690.

Abu Shawar and ES Atwell. 2004. An Arabic chatbot giving answers from the Qur'an. In *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 197–202. ATALA.

Bayan Abu Shawar. 2011. A chatbot as a natural web interface to Arabic web qa. *iJET*, 6(1):37–43.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Richard Wallace. 2003. The elements of AIML style. *Alice AI Foundation*.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale Arabic error annotation: Guidelines and framework. In *LREC*, pages 2362–2369.