Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 117 (2017) 208–216

Procedia
Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

Building a First Language Model for Code-switch Arabic-English

Injy Hamed[a,*], Mohamed Elmahdy[a], Slim Abdennadher[a]

[a]*The German University in Cairo, New Cairo, Cairo Governorate, 11432, Egypt*

Abstract

The use of mixed languages in daily conversations, referred to as "code-switching", has become a common linguistic phenomenon among bilingual/multilingual communities. Code-switching involves the alternating use of distinct languages or "codes" at sentence boundaries or within the same sentence. With the rise of globalization, code-switching has become prevalent in daily conversations, especially among urban youth. This lead to an increasing demand on automatic speech recognition systems to be able to handle such mixed speech. In this paper, we present the first steps towards building a multilingual language model (LM) for code-switched Arabic-English. One of the main challenges faced when building a multilingual LM is the need of explicit mixed text corpus. Since code-switching is a behaviour used more commonly in spoken than written form, text corpora with code-switching are usually scarce. Therefore, the first aim of this paper is to introduce a code-switch Arabic-English text corpus that is collected by automatically downloading relevant documents from the web. The text is then extracted from the documents and processed to be useable by NLP tasks. For language modeling, a baseline LM was built from existing monolingual corpora. The baseline LM gave a perplexity of 11841.9 and Out-of-Vocabulary (OOV) rate of 4.07%. The gathered code-switch Arabic-English corpus, along with the existing monolingual corpora were then used to construct several LMs. The best LM achieved a great improvement over the baseline LM, with a perplexity of 275.41 and an OOV rate of 0.71%.

Keywords: Automatic Speech Recognition; language model; code-mixing; code-switching; Arabic-English corpus; web corpus; web crawling

1. Introduction

The Arabic language is one of the most popular languages in the world. According to the 2017 edition of Ethnologue [7], the Arabic language has around 290 million native speakers (ranking 4*th* among all languages) and a total of around 422 million total speakers (ranking 5*th* among all languages). There are three types of the Arabic language: classical Arabic, modern standard Arabic (MSA), and dialectal/colloquial Arabic. The classical Arabic is the standard and most formal type of Arabic, which is used in the Quran and early Islamic literature. MSA is a direct descendant of the Classical Arabic and is the official form of Arabic taught in schools. MSA is a simplified version of Classical Arabic, having some simplifications such as the removal of diacritic marks. MSA is used in formal contexts such as formal writings and speeches, news broadcasts, movies' subtitling and education. Every Arabic-speaking country has its own dialect, and possibly sub-dialects. Dialects can differ radically from MSA, a linguistic characteristic referred to as "Diglossia", and are thus considered by researchers to be a separate language

[*] Corresponding author. Tel.: +2-010-0053-4848.
E-mail address: injy.hamed@guc.edu.eg

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2017.10.111

[27]. Therefore, MSA is considered as a second language for all Arabic-speakers and serves as a lingua franca across the Arabic-specking countries. Dialectal (Colloquial) Arabic is the language used in everyday conversations as well as informal writings such as blogs, chats and comments on on-line social media. Unlike MSA, dialectal Arabic does not necessarily have a standard written form.

Besides the three forms of Arabic, it has become a common trend by many Arabic-speakers to use more than one language in daily conversations. This phenomenon of bilingualism/multilingualism, referred to as "code-switching", has become of interest for linguists as well as researchers in the fields of language technologies. Several factors have given rise to this phenomenon including globalization, urbanization, immigration and international businesses and communication. Colonization has also played a role in the introduction of this phenomenon in Arab countries. Mixed speech is seen in several Arab countries, where for example, English is commonly used in Egypt and French in Morocco, Tunisia, Lebanon and Jordon.

Code-switching is defined as the fusion of two distinct languages: primary/matrix language (which is spoken in majority) and secondary/embedded language (by which words or phrases are embedded into the conversation). There are two types of code-switching:

- Inter-sentential: defined as switching languages from one sentence to another. For example, "في النهاية يكون الشكل كما يلي. Please make sure you have a correct output." (In the end it should look as follows. Please make sure you have a correct output.)
- Intra-sentential (also known as "code-mixing"): defined as using multiple languages within the same sentence. For example," غالبا ما تباع الcase بالpower supply الخاص بها." (Usually the case is sold with it's own power supply.)

In the scope of this paper, we will use the term "code-switching" to refer to both types of code-switching.

Code-switching has become a prevalent phenomenon among several bilingual/multilingual societies, such as Cantonese-English in Hong Kong [8], Mandarin-English in Singapore and Malaysia [9], Spanish-English in Hispanic communities in the United States [10], Turkish-German in Germany [11] as well as Italian-French and German-Italian in Switzerland [12]. Consequently, there has been a rising demand on multilingual automatic speech recognition (ASR) systems to be able to recognize such mixed speech. There are generally two main approaches to build a multilingual ASR. The first approach (language-dependent) is a multi-pass approach which involves a language boundary detection (LBD) algorithm to determine where language switch occurs. The LBD algorithm divides the input utterance into segments that are language-homogeneous. The language identity of each segment is identified using a language identity detection (LID) algorithm. The corresponding language-dependent automatic speech recognition system (ASR) is then used. The second approach (language-independent) is considered to be a more holistic way to build a multilingual ASR. It involves building an acoustic model, a language model and a pronunciation dictionary that encompass the languages in the mixed language speech. Recognition is then done in a one-pass approach.

Usually, the performance of language-independent systems is lower than that of language-dependent systems [4, 5]. However, this loss in accuracy is compensated by several advantages. The language-independent approach is generally a Simpler and faster approach for building multilingual ASRs. Language-independent multilingual ASRs do not rely on the existence of fast and reliable LBD and LID algorithms, which are still challenging tasks. In language-dependent ASRs, the final performance depends on the performance of the LBD and LID algorithms as well as the monolingual ASRs. A poor performance in any of the three blocks affects the overall system performance. Moreover, language-independent approaches are more suited for building multiligual ASRs that handle code-mixed speech, as the language-homogeneous segments will be very short and the performance of the LID algorithm will directly affect the speech recognition accuracy.