Code-Switching in Arabic Dialect Corpora: Compression vs Traditional Machine Learning Classifiers to Detect Code-switching

TAGHREED TARMOM

School of Computing, University of Leeds

sctat@leeds.ac.uk

t.a.tarmom@hotmail.com

WILLIAM TEAHAN

School of Computer Science, Bangor University

w.j.teahan@bangor.ac.uk

ERIC ATWELL

School of Computing, University of Leeds

E.S.Atwell@leeds.ac.uk

MOHAMMAD ALSALKA

School of Computing, University of Leeds

M.A.Alsalka@leeds.ac.uk

Abstract

The occurrence of code-switching in online communication when a writer switches between multiple languages presents a challenge for natural language processing (NLP) tools, since they have been designed for texts written in one language. There have been relatively few studies concerning the automatic detection of code-switching in Arabic texts from social media platforms such as Facebook. A contributory factor is that dialect identification for Arabic has been found to be more difficult than for other languages. Consequently, this paper presents detailed

research on ways for the automatic detection of code-switching in Arabic text. It adopts the compression-based toolkit (Tawa) (Teahan, 2018) and the Waikato Environment for Knowledge Analysis (Weka) data analytic tool (Hall et al., 2009) for the detection of code-switching in Arabic written text, specifically on text taken from Facebook.

In this paper, several new Arabic corpora are introduced. The first corpus is a new code-switching corpus that contain samples of Arabic code-switching for evaluating a compression- based approach and traditional machine learning classifiers to the automatic detection of code-switching in Arabic text. This is the Bangor Arabic–English Code-switching corpus (BAEC). To our knowledge, there are not any other available Arabic code-switching corpora from Facebook. Thus, it was necessary to build a new corpus for our research. The BAEC consists of 45,251 words and is 436 KB in size (see Figure 1). Two Saudi university researchers that have extensive knowledge in Egyptian dialect verified the quality of the BAEC tagging. If they disagreed on a particular word, whether it was in MSA or in an Arabic dialect, they looked at the Arabic online dictionary (www.almaany.com) and came to an agreement together.

The second corpora are new non-code-switching corpora for training language models. These are the Saudi Dialect Corpus (SDC), the Egyptian Dialect Corpus (EDC). A 210,396-word corpus called the Saudi Dialect Corpus (SDC) was built for training the Saudi model. The SDC corpus was built to contain the mixed dialects of Saudi Arabia. It was collected from social media platforms, such as Facebook and Twitter, and is 2,018 KB in size (see Figure 2). The Egyptian dialect corpus (EDC) that we constructed consists of 218,149 words and is 2,024 KB in size. It was also collected from the social media platform Facebook (see Figure 3). These new corpora are different from Arabic dialect corpora presented recently at LREC2018 in both data type and purpose. For example, Alshutayri and Atwell (2018) built the GLF corpus which contains all Arabic dialects in the Gulf region while the SDC corpus contains just Saudi dialects. Also, they built the EGY corpus which has Egyptian and Sudanese dialects. In contrast, the EDC has just the Egyptian dialect. All these new corpora will soon be made publicly.

```
<example id="137">
  <text>
    <MSA>الدرس الثالث من دروس المستوى الأول في اللغة الانجليزية المُقدم من</MSA>
    <English>iCareer</English><Egypt>المُستوى ما تنسوش أنه بعد نهاية .</Egypt>
    <MSA>سيتم فتح امتحان</MSA><English>online</English><MSA>مجاني</MSA>
    <English>For more listening:</English>
    <URL>www.rong-chang.com/easyspeak
     www.esl-lab.com</URL>
    <E.hashtag>#iCareer_English</E.hashtag>
  </text>
</example>
```

Figure 1: A sample from the BAEC Corpus.

ارسل لي وش تبي وابشر ولا عادي الوقت مفتوح
ادخل اسوي وللحين ماردو لهم يومين مين قد واج
أسوي كيف اقدر اجيب مو مرا قوية يعني يدوب
سيارة هذا الي صار معي ممكن لو سمحتي جربي الب

Figure 2: A sample from the SDC Corpus.

الناس بتشرب سجاير في الاسانسير ومفيش اي مراعاة للناس
برده لما نقعد في حتة والدخان كله يجي في وشنا ويقولك ا

Figure 3: A sample from the EDC Corpus.

Three experiments were performed as part of the evaluation of the compression-based approach (provided by Tawa) and traditional machine learning classifiers such as the Sequential Minimal Optimization (SMO) classifier (provided by Weka) to detect code-switching in Arbic Facebook text. These were to: (1) detect code-switching between the Egyptian dialect and English, (2) detect code-switching between the Egyptian dialect, the Saudi dialect and English and (3) detect code-switching between the Egyptian dialect, the Saudi dialect, MSA and English. Our experiments firstly showed that Tawa achieved a higher accuracy rate than Weka when the training corpus correctly represents the language or dialect under study. So, when this condition is satisfied the compression-based approach will be a more effective approach for automatically detecting code-switching in written Arabic text. Secondly, when Arabic and English are classified using Weka, the CharacterNGramTokenize feature is a more appropriate feature to use than the WordTokenizer feature since the difference between these two languages is best modelled using characters. Thirdly, the CharacterNGramTokenize feature is also a more appropriate feature to compare between Weka and Tawa since Tawa is a character-based model.

The first experiment focused on the detection of code-switching between the Egyptian dialect and English. Tawa obtained an accuracy of 99.8% on testing data from the BAEC which is 2.3% higher than the SMO classifier. The second experiment investigated the automatic detection of code-switching between the Egyptian dialect, the Saudi dialect and English. Tawa achieved an accuracy of 97.8% which is 17.1% higher than the SMO classifier. Finally, the third experiment detected code-switching between the Egyptian dialect, the Saudi dialect, MSA and English. The SMO classifier obtained an accuracy of 60.2% which is 6.9% higher than Tawa.

Clearly, the MSA corpus used to train the MSA PPM model in the third experiment did not represent MSA text in Facebook, since it was built from news websites. As part of future work, a possible solution to overcome this issue is to build a new MSA Facebook corpus trained on MSA text specially taken from Facebook. In addition, the distinction between MSA and Arabic dialects is a very difficult task because most of the Arabic users, especially Saudis and Egyptians, mix MSA with their dialects. Also, using a new English corpus containing all the possible abbreviations should improve the results further. In fact, the detection of TEng (some English words were written using Arabic letters) and MAE (mixing Arabic with English words to produce one mixed word) words provides one of the biggest challenges for NLP tools, since it was an unforeseen issue and we did not have enough training data to provide an effective means for identifying these phenomena.

References

Alshutayri, A., & Atwell, E. (2018, May). Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers. In *OSACT 3: The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools* (p. 54).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations* newsletter, 11(1), pp.10-18.

Teahan, W. (2018). 'A compression Based Toolkit for Modelling and Processing Natural Language Text'. *Journal of Information. MDPI Publishers*. Accepted for publication.