Filtering Dialectal Arabic Text in Two Large Scale Annotation Projects

Wajdi Zaghouani,[1] Nizar Habash,[3] Houda Bouamor,[1] Ossama Obeid,[1]
Sawsan Alqahtani,[2] Mona Diab,[2] and Kemal Oflazer[1]
[1]Carnegie Mellon University in Qatar [2]George Washington University [3]New York University Abu Dhabi

Dialectal Arabic (DA) refers to the different dialects of Arabic spoken in the Arab world. Although there are many common shared features between DA and Modern Standard Arabic (MSA), DA differs substantially in all linguistics aspects such as syntax, semantics and more specifically morphology. Typically, native speakers of Arabic mix DA and MSA in the same conversation or even the same sentence. This phenomenon is defined as code switching. Thus, written DA texts, such as online user-generated content are generally noisy. For instance, the online AlJazeera user comments corpus contains, in addition to MSA words, many dialectal words. This corpus was created within the Qatar Arabic Language Bank annotation project (QALB), a large-scale error correction and annotation effort that aims to create a manually corrected corpus of errors for a MSA texts (Zaghouani et al. 2014). In this abstract, we briefly describe how we handled and corrected words mistakenly written in DA in AlJazira user comments. We followed two strategies included in two large scale annotation projects, namely the QALB project the Optimal Diacritization (OptDiac) project.

In comparison to MSA, where there are clear spelling standards and conventions, DA do not have official orthographic standards since they were not commonly written until recently. Today, Arabic dialects are often seen in social media, such as Facebook, Twitter or online comments. In the QALB project we created AlJazeera online user comments corpus. In this project, we aimed at reducing the various spelling inconsistencies that frequently occur, and asked the annotators to flag the highly dialectal cases using the QAWI tool (Obeid et al. 2013). The guidelines created, classify dialectal word errors into five categories inspired by Habash et al. (2008): (i) dialectal lexical choice, (ii) pseudo-dialectal lexical choice, (iii) morphological choice, (vi) phonological choice and (v) closed class dialectal words. Only the last three categories are considered for correction while the first two are flagged and marked as DA by the annotator. Therefore, we neither address DA spelling normalization, nor do we systematically translate dialectal words into MSA as we recognize that the Arabic language is in a diglossic situation and borrowing is frequent.

Within the OptDiac project framework, the main goal is to manually create a large-scale annotated corpus with diacritics for a variety of MSA texts from various sources and covering more than ten genres. Our DA study corpus covers mainly online user comments. In order to ensure that DA sentences are discarded from our corpus, we proceed as follows. In a first pass, we filter out the sentences tagged as Dialectal Arabic (DA) using AIDA, an automatic identification of DAT tool (Elfardy et al. 2014). AIDA categorizes token level words and sentences as MSA or DA using a set of language models, dictionaries, a morphological analyzer and phonological change rules. Once sentences identified as dialectal are removed, we proceed with the annotation using MANDIAC, a web-based annotation tool and a work-flow management interface (Obeid et al., 2016). In the second pass, annotators manually filter out the DA sentences that were not detected during the initial pass. For this, we provide thhem with a Flag button to tag the DA words. Using the pipeline described above, we were able to filter out all the DA sentences in the corpus and keep only the MSA ones. To sum up, filtering out and processing the DA is a fairly complex task and a manual human annotation pass using expert annotators is strongly recommended.

Acknowledgements

This publication is made possible by grants NPRP-6-1020-1-199 and NPRP-4-1058-1-168 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. AIDA: Identifying Code Switching in Informal Arabic Text. InProceedings of the Workshop on Computational Approaches to Linguistic Code Switching (EMNLP 2014)

Ossama Obeid, Houda Bouamor, Wajdi Zaghouani, Mahmoud Ghoneim, Abdelati Hawwari, Sawsan Alqahtani, Mona Diab, Kemal Oflazer. MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization. In Proceedings of the 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media (LREC2016)

Ossama Obeid, Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Kemal Oflazer, and Nadi Tomeh. 2013. A Web-based Annotation Framework For Large-Scale Text Correction. In Proceedings of IJCNLP 2013

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for Annotation of Arabic Dialectness. In Proceedings of the Workshop on HLT & NLP within the Arabic world (LREC 2008)

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 Proceedings of the LREC 2014