Discovering Orthographic and Morphological Variances in Low-Resourced Languages

Anonymous ACL submission

It is evident that the social media has become a space for unstructured languages, even more, languages develop or form on the social media. This research draws attention to a major NLP challenge, lexical coverage in low-resourced morphologically-rich and Latinised languages. It focuses on Arabizi, a widely used variety of Arabic, hence naturally rich in morphology, but written in Latinscript, thus naturally lacking a unified orthography as well (Yaghan, 2008).

Latinised languages are born when bilinguals express their non-Latin language in Latinscript. This is common for Arabic, Greek, Farsi, Hindi, Filipino, Mandarin, and other Far Eastern languages. The rationale behind this is either the lack or difficulty of using non-Latin script keyboards. Studies have shown that Arabizi makes a 12% of the Latin script tweets in Lebanon and 25% of the Latin script tweets in Egypt (Tobaili, 2016), it is a common way of communication among the youth (Keong et al., 2015; Muhammed et al., 2011; Allehaiby, 2013) and proven to be a key communication medium in relevant events in the Arab world such as the Arab spring (Basis-Technology, 2012) yet it had been filtered out in several researches (Duwairi and Qarqaz, 2014; Al-Kabi et al., 2014, 2013) missing on relevant information from a considerable portion of the population.

Aiming to make artificial sense out of Arabizi, we propose to analyse sentiment from this overlooked low-resourced language. We started by creating a sentiment lexicon (SenZi) through several stages of automatic translation and manual transliteration reaching 607 positive and 1,383 negative words. We collected, preprocessed, annotated, and balanced a dataset of 800 positive and 800 negative tweets to evaluate SenZi using a simple score aggregation lexicon-based approach[1] achieving an F1-score of 0.57.

We analysed the errors to find that the majority of the missed sentiment words are either orthographic or morphological forms of the words in SenZi. Therefore, creating a sentiment lexicon with one form of each sentiment word for morphologically-rich and Latinised languages is insufficient to cover the number of inflections and spellings for each word. For example, the Arabizi word *7ob* meaning *love* produces at least 100 inflectional forms multiplied by the number of different spellings for each form.

We addressed this challenge by harvesting a corpus of 1M Facebook Arabizi comments and projected it into a vector space of word embeddings using the fastText skip-gram model (Bojanowski et al., 2016). We discovered the different orthographic and inflectional forms of each sentiment word by retrieving its nearest word neighbours. We identified these forms by measuring the similarity of the consonant-letter-sequence with the SenZi words. The result was a new publicly available Arabizi sentiment lexicon consisting of 11.3K positive and 13.3K negative words pushing our baseline F1-score by a significant 15%.

Table 1: Evaluation of SenZi

	Recall	Precision	F1-Score	Accuracy
SenZi	0.56	0.59	0.57	0.58
SenZi-Embed	0.79	0.66	0.72	0.69

Word embeddings proved to be an excellent technique to leverage morphologically-rich and Latinised languages easily. Next, we will explore cross lingual word embeddings (Glavas et al., 2019; Ruder et al., 2017) to discover more information and transliterate among low-resourced languages that are transcribed in different scripts such as Arabic and Arabizi, Greek and Greeklish, Farsi and Finglish, and Hindi and Hinglish.

[1] Since this is a 2-class classification on a balanced dataset, we randomised a sentiment class for tweets that scored 0.

References

Mohammed Al-Kabi, Amal Gigieh, Izzat Alsmadi, Heider Wahsheh, and Mohamad Haidar. 2013. An opinion analysis tool for colloquial and standard arabic. In *The Fourth International Conference on Information and Communication Systems (ICICS 2013)*, pages 23–25.

Mohammed N Al-Kabi, Amal H Gigieh, Izzat M Alsmadi, Heider A Wahsheh, and Mohamad M Haidar. 2014. Opinion mining and analysis for arabic language. *International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher*, 5(5).

Wid H Allehaiby. 2013. Arabizi: An analysis of the romanization of the arabic script from a sociolinguistic perspective. *Arab World English Journal*, 4(3):52–62.

Basis-Technology. 2012. The burgeoning challenge of deciphering arabic chat.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Rehab M Duwairi and Islam Qarqaz. 2014. Arabic sentiment analysis using supervised classification. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pages 579–583. IEEE.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Yuen Chee Keong, Othman Rahsid Hameed, and Imad Amer Abdulbaqi. 2015. The use of arabizi in english texting by arab postgraduate students at UKM. *The English Literature Journal*, 2(2):281–288.

Randa Muhammed, Mona Farrag, Nariman Elshamly, and Nady Abdel-Ghaffar. 2011. Summary of arabizi or romanization: The dilemma of writing arabic texts. In *Jīl Jadīd Conference, University of Texas at Austin*, pages 18–19.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

Taha Tobaili. 2016. Arabizi identification in twitter data. *ACL 2016*, page 51.

Mohammad Ali Yaghan. 2008. arabizi: A contemporary style of arabic slang. *Design Issues*, 24(2):39–52.