Arabic Metaphor Corpus (AMC) with Semantic and Sentiment Annotation

Israa Alsiyat, Scott Piao and Mansour Almansour

School of Computing and Communications, Lancaster University, UK
{i.alsiyat,s.piao,m.s.almansour}@lancaster.ac.uk

Abstract

Metaphor is an inseparable part of communication in Arabic language. Over the past years, Arabic metaphors have been increasingly used in online communication, which has drawn attention of researchers. Arabic metaphors in online context have been evolving and show distinct features, reflecting the need for speedy online communication compared to traditional written forms. While Arabic metaphors have intrinsic ambiguity of word sense, the metaphors used online present an even higher challenges for accurate interpretation of their true meanings in their contexts. For example, no reliable resources can be found for online variants of Arabic metaphors for interpretation, newly emerging metaphor structures are not registered in standard Arabic literature, they can have different writing styles in different Arabic dialects, and frequently appearing typos etc. Therefore, we need a corpus that reflects online Arabic metaphor usage annotated with their key features for the investigation of Arabic metaphor usage in online context and to develop tools for identifying and classifying them. However, as far as we are aware, there is no such corpus publicly available yet.

To address this issue, we have built an Arabic Metaphor Corpus (AMC), which contains 1,000 online book reviews from LARB Corpus (Aly and Atiya, 2013). The AMC was manually annotated by two native Arabic annotators for sentiment of reviews and metaphors, then we merge different sets of annotation into gold standard annotations. Furthermore, the reviews were tagged using the UCREL Arabic Semantic Tagger AraSAS (El-Haj *et al*., 2022) to facilitate detailed analysis of semantic patterns and contexts of the metaphors.

The AMC is available online (https://github.com/IsraaMousa/IsraaArabicMetaphor/find/main). The AMC corpus is a unique resource for the study of Arabic metaphors in online review context, and will contribute a valuable corpus resource to the corpus linguistics community. The average length of the reviews is 107.968 tokens, and the longest and shortest reviews have 1,084 and 2 tokens respectively. In the AMC, according to our gold annotation, there are 702 and 171 positive and negative reviews, making up 70.2% and 17.1% of the reviews respectively. The remaining 127 reviews are judged to be neutral. In regard to the metaphors annotated in the AMC, 602 and 272 metaphors are manually classified as positive and negative ones respectively. The remaining 126 metaphors are classified as neutral.

In this presentation (see the presentation slides), we report the data selection criteria, methodology of manual annotation, and statistical information of data structure and annotation. We will also discuss challenging issues arisen during the corpus construction process.

Keywords: Arabic Metaphor Corpus, Sentiment Analysis, Metaphor, Arabic Corpus Linguistics, Arabic NLP

References:

Aly, M. and A. Atiya (2013). Labr: A large scale arabic book reviews dataset. In Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria. Volume 2: Short Papers. pp. 494–498.
LABR corpus is available at website:
https://scholar.googleusercontent.com/scholar?q=cache:0S-Kb9VPnocJ:scholar.google.com/+large+Arabic+book+reviews&hl=en&as_sdt=0,5 (accessed May 29, 2023).

El-Haj, M., E. de Souza, N. Khallaf, P. Rayson, and N. Habash (2022). AraSAS: The Open Source Arabic Semantic Tagger. In Proceedings of The 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection. Marseille, France, pp. 23–31.