Identification of Arabic dialects and disambiguation of language for natural language processing applications

Technology #cu14014

Although natural language acquisition programs have been developed for a myriad of languages, they have only tenuously been applied to Arabic, due to ambiguities and complexities in the way it is written and the large number of regional dialects commonly used in informal communication. This technology describes two software packages, MADAMIRA (Morphological Analysis and Disambiguation for Dialectal Arabic) and AIDA (Automatic Identification of Arabic Dialectals) that, working in tandem, can perform linguistic analyses on raw Arabic text required for natural language processing and additionally assign dialects on both sentence and token levels while generating an English translation. Utilized together, these technologies may be used to identify and translate a variety of written or spoken Arabic dialects into English for commercial and educational purposes.

A flexible software package that extracts and clarifies the linguistic information needed for accurate Arabic language processing

Developed in conjunction, the MADAMIRA and AIDA software suites allow for the generation of linguistic information such as tokenization, diacritization, and part-of-speech tagging from written Arabic words in a variety of dialects. Leveraging machine-learning algorithms and established morphological analyzers, ambiguity is reduced while limiting bias. The accuracy of these programs has been demonstrated at the Center for Computational Learning Systems at Columbia University. Furthermore, because AIDA is Java-based, it can be configured as either a web-based interface or packaged for offline processing, including in mobile devices. Together, these software packages allow for the rapid translation of various Arabic dialects into English with high accuracy and a low level of ambiguity.

Lead Inventor:
Owen Rambow, Ph.D.

Applications:
• Text-to-speech
• Named entity recognition
• Automatic speech recognition
• Identification and translation of Arabic dialects for educational and commercial applications
• Development of a dialectal Arabic dictionary
• Development of voice recognition translation device of spoken Arabic

Advantages:

• Single software package capable of performing several natural language processing tasks
• Unbiased natural language processing for Arabic, providing flexibility for developers building applications requiring NLP
• High accuracy in predicting linguistic features such as part-of-speech, lemmatization, diacritization, and tokenization
• Output can be configured to the user’s preference
• Includes Egyptian, Iraqi, Levantine, and Moroccan dialects
• Can be used on any computing platform capable of running Java, including mobile devices
• Can be used as a stand-alone program or in a client-server mode

Tech Ventures Reference: IR CU14012, CU14014

Related Publications:


Inventors

Owen Rambow