Arabic Language Disambiguation for Natural Language Processing Applications

Processing the Arabic language into its elementary linguistic components is particularly challenging due to two factors: the morphology of Arabic (how Arabic words are put together) is complex, and Arabic orthography (how Arabic is written) is highly ambiguous, since Arabic is typically spelled without short vowels and other diacritical markers. Additionally, Arabic has many dialects that vary from Standard Arabic (formal language used in education and mainstream print media) in terms of morphology and lexicon, and have no standard writing systems. MADAMIRA, is a software suite for morphological analysis and disambiguation of Arabic and its dialects. MADAMIRA can perform several types of linguistic analyses on raw Arabic text as required for natural language processing (NLP). The MADAMIRA toolkit uses machine-learning algorithms and established Arabic morphological analyzers to report the linguistic features of each word in context. Downstream NLP products and tools may then use the analyses yielded from MADAMIRA for further work.

The MADAMIRA software suite extracts and clarifies the linguistic information needed to support accurate Arabic natural language processing

MADAMIRA provides linguistic information such as tokenization, diacritization, lemmatization, part-of-speech tagging, full morphological tagging, base phrase chunking and named entity recognition for each Arabic word received as input. MADAMIRA users can then use this information to create the analysis best suited for their application. The high accuracy of MADAMIRA in correctly predicting linguistic features of Arabic words has been demonstrated at the Center for Computational Learning Systems at Columbia University.

Lead Inventors (alphabetical order):

- Mona Diab, Ph.D.
- Nizar Habash, Ph.D.
- Owen Rambow, Ph.D.

Applications Provided:

- Tokenization
- Part-of-Speech tagging
- Morphological disambiguation for full range of morphological features
- Lemmatization
- Diacritization
- Named entity recognition
- Base phrase chunking

**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, lemmas, diacritics, and tokenization

**Patent information:**

Patent Pending

**Licensing Status:**

Available for licensing and sponsored research support

Tech Ventures Reference: IR CU14012

**Selected Related Publications:**


- [M. Diab. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking. MEDAR 2nd International Conference on Arabic Language Resources and Tools, April, Cairo, Egypt, 2009](#)

- [N. Habash, O. Rambow, and R. Roth. MADA+TOKAN, A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, 2009](#)


Inventors

Mona T. Diab