AIDA: Automatic Identification and Glossing of Dialectal Arabic

Heba Elfardy and Mona Diab
Center for Computational Learning Systems
Columbia University
{heba,mdiab}@ccls.columbia.edu
http://nlp.ldeo.columbia.edu/aida/

Description

AIDA is a system for dialect identification, classification and glossing on the token and sentence level for written Arabic. Automatic dialect identification in Arabic is quite challenging because of the diglossic nature of the language and informality associated with the typical genres where dialectal Arabic (DA) is used. Moreover, DA lacks a standard orthography. Additionally the abundance of faux amis between the different varieties of Arabic, namely between Modern Standard Arabic (MSA) and DA, exacerbates the challenge of identifying dialectal variants. Hence identifying whether a (sequence of) token(s) is MSA or DA and providing an MSA-Gloss for the dialectal tokens in an utterance can aid Arabic MT in handling such informal genres more accurately.

AIDA aggregates several components including dictionaries and language models in order to perform named entity recognition, dialect identification & classification and MSA & English linearized glossing of the input text. The default output produces the following information for each token in the input text:

1. CLASS: this field displays whether a given word is DA, MSA or unknown (MSA, DA or UNK), and for dialectal words it identifies the class: either Egyptian, or Other (another Arabic dialect);

2. NE: Whether the word is a named-entity (NE) or not and if it is a NE, the NE class it belongs to (Person, Organization, GeoPolitical entity, Location);

3. MSA-Gloss: For dialectal tokens this field displays the MSA equivalents of a token ordered by their frequency of occurrence in Arabic Gigaword;[1]

4. English-Gloss: The English equivalents of the given token.

Ex:

Input (UTF8)	ده	اللي	بيحصل	في	مصر	في	الوقت	الراهن
Input(BW)[2]	dh	El~y	byHSl	fy	mSr	fy	Alwqt	AlrAhn
Class	DA	DA	DA	DA/MSA	-	DA/MSA	MSA	MSA
NE:	-	-	-	-	GPE	-	-	-
MSA-Gloss:	*lk	Al*y	yHdv	fy	mSr	fy	Alwqt	AlrAhn
ENG-Gloss:	that	what	happen	in	Egypt	in	the time	the current

The output is configurable allowing the user to choose the output encoding as well as the user's preference on what tagging information to display. AIDA is accessible through a configurable web-based interface as well as a packaged pipeline that is available for offline processing.

[1] http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2009T30
[2] We use Buckwalter transliteration scheme: http://www.qamus.org/transliteration.htm