



Unsupervised Induction of Modern Standard Arabic Verb Classes Using Syntactic Frames and LSA

Neal Snider (Stanford University), Mona Diab (Columbia University)



Abstract

We exploit the resources in the Arabic Treebank (ATB) and Arabic Gigaword (AG) to determine the best features for the novel task of automatically creating lexical semantic verb classes for Modern Standard Arabic (MSA). The verbs are classified into groups that share semantic elements of meaning as they exhibit similar syntactic behavior. The results of the clustering experiments are compared with a gold standard set of classes, which is approximated by using the noisy English translations provided in the ATB to create Levin-like classes for MSA. The quality of the clusters is found to be sensitive to the inclusion of syntactic frames, LSA vectors, morphological pattern, and subject animacy. The best set of parameters yields an $F_{\beta=1}$ score of 0.456, compared to a random baseline of an $F_{\beta=1}$ score of 0.205.

Objective

- Automatically induce MSA verb classes on a large scale
- Classes defined by similar event structures
- Using syntactic alternation behavior (Levin 1993)
- Test efficacy of general features and those specific to Arabic

Methods

- Data
 - Arabic Treebanks 1, 2, 3 (LDC 2003)
 - 800,000 Arabic tokens
 - Manually parsed and lemmatized
 - Arabic Gigaword 2
 - Morphologically disambiguated (MADA, Habash & Rambow, 2005)
- Cluster using Fuzzy clustering (R statistical package)

Syntactic Frames

Sisters to V in a VP constituent

- NP arguments (NP-SBJ, NP-OBJ, etc.)
- PPs deemed essential to verb meaning annotated PP-CLR
- SBAR
- NP-TPC due to subject extraction in SVO configurations

LSA Similarity

Similarity vectors of verbs

- derived from Latent Semantic Analysis of Arabic Gigaword
- Dimension reduction by Singular Value Decomposition

Features

Verb	Morphology		Subject Animacy			Syntactic Frames			LSA Similarity		
	Root	Pattern	Pro-dropped	Pronoun	Prop Name	<NP-SBJ NP-OBJ>	...	<NP-SBJ PP-CLR- علي >	1	...	44
اغار	gAr	4	0.43	0.28	0.11	0.3	...	0.5	0.12	...	0.52
...
تمزق	mzq	6	0.34	0.26	0.45	0.4	...	0	0.23	...	0.01

Morphology

- Templatic verbal morphology

Root:

k t b

↓ ↓ ↓

Pattern 1: 1V2V3 => kataba (write)

Vocalization: a

- 10 patterns in corpus

Subject Animacy

- Use proxy features

- Count frequency the verb's subject is:

-Pro-dropped اغتسله (He washed it)

-Pronoun هو يصعد (He is climbing)

-Proper name بوش ينام (Bush is sleeping)

Results

Evaluation

- Derived 184 Gold standard Levin-like classes from noisy English translations of 406 verb types in ATB
- Cluster overlap metric (Chklovski & Mihalcea, 03)
- F-score, combining precision and recall

Example Clusters

- Convening verbs (good prec & recall)
 - >aloqaY [meet], \$ahid [view], >ajoraY [run an interview], {isotaqobal [receive a guest], Eaqad[hold a conference], >aSodar [issue].
- Say verbs (low recall)
 - *akar [mention], >afAd [report]
 - *GOLD: >aEolan [announce], >a\$Ar [indicate], *akar [mention], >afAd [report], SaraH [report/confirm], \$ahid [relay/witness], ka\$af [uncover]
- Occurrence verbs
 - Eamil [work continuously on], jA' [occur], {isotamar [continue], mA zAl [remain], baqiy[remain], jaraY [occur]
 - *GOLD: jA'[occur], HaSal [happen], jaraY [occur]

Feature Evaluation

Syntactic frames (+)	p<.03
Subject animacy (+)	p<.002
Morphological pattern (-)	p<.001
Root	n.s.
LSA (+)	p<.001

	F-Score	F-Score Prec
Frames	38%	50%
+SubjAnimacy		
Frames	46%	68%
+SubjAnimacy		
+LSA		
Random (baseline)	21%	37%