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-value of 0.456, compared to a random baseline of an F-value 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)
  • Gigaword 2
  • Manually parsed and lemmatized
• Arabic Gigaword
  • Morphologically disambiguated (MADA, Habash & Rambow, 2005)
• Cluster using Fuzzy clustering (R statistical package)

Features
• Morphology
  • Templatice verbal morphology
  • Root:
    • 1/2/3
• 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

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

Subject Animacy
• Use proxy features
  • Count frequency the verb’s subject is:
    – Pro-dropped
    – Pronoun
    – Proper name

Evaluation
• Derived 184 Gold standard Levin-like classes from noisy English translations of 406 verb types in ATB
• Cluster overlap metric (Chkovski & Mihalcea, 05)
• F-score, combining precision and recall

Results

<table>
<thead>
<tr>
<th>Morphology</th>
<th>Subject Animacy</th>
<th>Syntactic Frames</th>
<th>LSA Similarity</th>
</tr>
</thead>
</table>
| Verb       | Root | Pattern | Pro-dropped | Pronoun | Prop Name | <NP-SBJ PP-CLR> | <NP-SBJ PP-CLR> | 1 | ...
| ánagar     | gAr  | 4      | 0.43       | 0.28    | 0.11      | 0.3            | 0.5            | 0.12 | ...
| تمرة       | mzq  | 6      | 0.34       | 0.26    | 0.45      | 0.4            | 0            | 0.23  | ...


Example Clusters
• Convening verbs (good prec & recall)
  • >الوقاية [meet], >الشيد [view], >الأجر [run an interview], >الوقاية [receive a guest], >الأجر [hold a conference], >الأجر [issue]
• Say verbs (low recall)
  • وقأ [mention], >الباق [report]
  • GOLD: >الوقاية [announce], >الأجر [inform], >الوقاية [report], >الأجر [confirm], >الأجر [relay/witness], >الوقاية [mention]
• Occurrence verbs
  • >الوقاية [work continuously on], >أنا [occur], >الوقاية [continue], >الوقاية [remain], >الوقاية [remain], >الوقاية [occur]
  • GOLD: >الوقاية [occur], >الوقاية [happen], >الوقاية [occur]

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

Frames
+SubjAnimacy
Frames
+SubjAnimacy +LSA
Random (baseline)

F-Score F-Score Prec
38% 50%
46% 68%
21% 37%