Using Web Mining Techniques to Build a Multi-Dialect Lexicon of Arabic

Rania Al-Sabbagh†
Roxana Girju †
Mark Hasegawa-Johnson †
Abbas Benmamoun †
Rehab Duwairi‡
Eiman Mustafawi‡ ‡

†Department of Linguistics, University of Illinois at Urbana-Champaign, USA
‡ Department of Computer Science and Engineering, Qatar University, Qatar.
‡‡ College of Arts and Sciences, Qatar University, Qatar
What?

• A multi-dialect lexicon of Arabic: MSA, Moroccan, Egyptian and Gulf Arabic

• The entry in one dialect/variety (MSA)
  • mapped to its equivalents in the other dialects
  • tagged for POS, gender and number
  • identified for context (i.e. collocations)
Why?

- The lexicon is to be used for many NLP tasks like POS tagging, WSD and anaphora resolution among others.
- If we include MSA, we’ll cope with the tendency of leveraging tools and resources for Arabic dialects from the current ones available for MSA.
  
  http://www.ccls.columbia.edu/project/cadim-columbia-arabic-dialect-modeling
- Research on Arabic dialects is still in its infancy and lacks basic language processing tools and resources.
Previous Work [1]

- **Duh and Kirchhoff (2005)**
  - a POS tagger for ECA based on an initial ECA lexicon developed by using Buckwalter’s AraMorph (2002) that was originally developed for MSA.

  - Buckwalter’s system provided possible POS tags for 62% with accuracy rate of 62.76%.

  - Implications: considerable common vocabulary between the two varieties, we can expect considerable common contexts as well.
Rambow et al. (2005)

- used Rapp’s (1999) algorithm for inducing lexical mappings from unrelated corpora based on the correlation of co-occurrence patterns of words in both corpora.

- Using the city block distance as their similarity measure and log likelihood as their collocational association measure.

- No specific results were given about the performance of using Rapp’s algorithm on Arabic dialects.
Previous Work [3]

- **Al-Sabbagh and Girju (2010)**
  - The paper is the premises of the present study although the present paper is more extended.
  - Using circular acquisition of collocations – same technique used here – they try to build a lexicon of Egyptian Arabic and MSA.
  - Their achieve a recall rate of 70% and a precision rate of 72%
Our Approach

- **Underlying concepts**
  - Words are defined in terms of their co-occurring words in local contexts (Rapp 1999)
  - Words of the same meaning share the same co-occurrence patterns (the associationist approach)
  - Words of the same meanings share the same semantic features of gender and number

- **Problem?**
  - Large Corpora

- **Web as Corpus?**
  - Partially handles the problem:
    - Web content of Arabic dialects is still relatively small, in comparison to the Web content of MSA
    - Indexing techniques of Web documents on meta-search engines are usually problematic (duplication & 1000 limit)
Solution

- Still using Rapp’s algorithm (1999) but try to overcome its bottleneck (i.e. scarce corpora)

Circular Acquisition

- word co-occurrences in the first dialect are acquired *first* and then they are *validated* as possible co-occurrences for the words of the second dialect

- Gives a better technique to search Web documents avoiding the aforementioned problems of search engines.
- Collocations of the first dialect are acquired
- Collocations of the second dialect are acquired separately
- There are differences and similarities between the acquired collocations
- So we get into mapping the similarities
Visualization [2]

- Acquire the collocations of the most frequent dialect (MSA)
- Validate whether they are possible collocations of the 2nd dialect validation
- Valid collocations for 1st & 2nd dialects are validated for the 3rd dialect
- Valid collocations for 1st, 2nd & validation 3rd dialects are validated as possible collocations for 4th dialect
Real Example

الحديث /alhadiːθ/ (speaking): (MSA)

الكلام /akalam/ (Egyptian)

حجي /hagi/ (Gulf)

هدر /hadr/ (Moroccan)

• Synonyms
• Expected common local contexts
• **But**
  • With Parallel Acquisition, no common local contexts (i.e. collocations) were found
  • With Circular Acquisition, many common co-occurrences were found
Statistical Measurements

- Collocational Association:
  - Measured according to Pointwise Mutual Information (PMI) defined as:

\[
SI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}.
\]
POS and Semantic Tagging

- Main assumption:
  - Synonymous words share the same POS and semantic features of gender and number

- MSA has POS taggers (Diab et al. 2002) and lexicons of semantic features (Elghamry et al. 2008)

- Once the dialect word is recognized as a synonym of a given MSA; it acquires the same POS and semantic tags
Evaluation [1]

- Baseline & Manual Evaluation

(1) AraMorph (Buckwalter 2002) is used to evaluate two aspects of the proposed algorithm, namely POS tagging and the semantic features of gender and number given that AraMorph does not give ECA-MSA synonym.
Evaluation [2]

(2) Manual Evaluation

- used to evaluate POS tagging, semantic features labeling and synonym mappings

- Two human raters are used with Kappa Coefficient as the measure for inter-rater agreement.
Evaluation Metrics

(1) **Recall** =
\[
\frac{\text{# of dialect words mapped to MSA words}}{\text{Total number of dialect words}}
\]

(2) **Precision** =
\[
\frac{\text{# of dialect words correctly mapped to MSA words}}{\text{# of dialect words mapped to MSA words}}
\]

(3) **F-Measure** = \[
\frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})}
\]
Results

- Preliminary results are **promising**,

- More corpora are still to be compiled given the scarcity of the Web content of some dialects esp. Moroccan.
Future Work

- First, only content words are considered here; yet function words are expected to be more complicated because they are not to be defined in terms co-occurring words but in terms of their grammatical behavior.
- Second, synonyms across different dialects are not necessarily one-to-one alignments.
- Finally, inter-dialectical words are still beyond the scope of this paper and thus they should be considered in further research.