Tweet Conversation Annotation Tool with a Focus on an Arabic Dialect, Moroccan Darija

Stephen Tratz †, Douglas Briesch †, Jamal Laoudi ‡, and Clare Voss †

Army Research Lab †

Advanced Resource Technologies INC ‡
Outline

• Dialect Annotation Tool (DATool)

• Classifier Integration

• Results visualization using GEPHI
DATool

• Automatically collects conversation tweets given a user profile

• Allows for manual and automatic annotation of Darija tweet conversations

• Saves annotation results in network formats for downstream discourse analysis
Notes on Moroccan Darija

Moroccan Darija (Darija):

• A variety of Arabic

• Darija is non standardized and spoken only

• Low online content but slowly emerging

• Appears in multiple scripts including Arabic (e.g. منزل) and Roman (e.g. Manzil), and sometimes digits (e.g. 7a9l)

• Often mixed with other languages (e.g. Modern Standard Arabic (MSA), French)
Dialect Annotation Tool and Features

Manual Annotation

- All Darija: Fully highlighted tweet
- Partial Darija: Red check mark
- Non Darija: No annotation

Automatic Classifier

Summary Statistics

Export Annotation files in GEXF for use by GEPHI, a network visualization tool
Dialect Annotation Tool and Features

Automatic Annotation

• Multiple classification models can be used
• Each language id is colored differently
• Provides confidence score for each tweet
Classifier Building and Enhancing

Training Data

- Darija Query Results by search Engine (Darija)
- Newsblog commentary (Darija)
- Jokes (Darija)
- Zaidan & Callison-Burch, (MSA, Gulf, Egyptian, Levantine)

Classifier

- Max Entropy Model
- LDA Model
- Additional LDA-derived Features
- LDA Input Features
- Training data-derived features
Classifier Evaluation

Test Set (Arabic script tweets)

<table>
<thead>
<tr>
<th></th>
<th>#non-Darija</th>
<th>#Darija</th>
<th>#Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>387</td>
<td>1,013</td>
<td>1,400</td>
</tr>
</tbody>
</table>

Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline training data-derived features</td>
<td>96.9</td>
<td>24.1</td>
</tr>
<tr>
<td>w. LDA-derived features</td>
<td>97.2</td>
<td>44.1</td>
</tr>
</tbody>
</table>
From DATool to GEPHI

Information Network

Social Network

Social Information Network

Social Information Network Plus
Conclusion

• Built a tool to:
  - Collect and annotate Darija conversation tweets manually and automatically
  - Export results for use with network visualization tools

• Developed a method to build low-resource language corpora from online resources
Future Work

• Expand classifier to include other languages
• Explore other network diagrams:
  - Predict code-switching
  - Provide deep linguistic analysis
• How will pre-classification of data impact human annotation time?
• When to adjust confidence threshold in DATool?
Date- August 9th
Time- 16:15 to 17:15

Tweet Conversation Annotation Tool
with a Focus on an Arabic Dialect, Moroccan Darija

Stephen Tratz†, Douglas Briesch†, Jamal Laouafi†, Clare Voss†
†Army Research Laboratory, Adelphi, MD | ‡Advanced Resource Technologies, Inc., Alexandria, VA

ABSTRACT

This paper presents the DATool, a graphical tool for annotating conversations consisting of short messages (i.e., tweets), and the results we obtain in using it to annotate tweets for Darija, an historically unwritten Arabic dialect spoken by millions but not taught in schools and lacking standardization and linguistic resources. With the DATool, a native-Darija speaker annotated hundreds of mixed-language and mixed-script conversations at approximately 250 tweets per hour. The resulting corpus was used in developing and evaluating Arabic dialect classifiers. The DATool supports downstream discourse analysis of annotated "conversations" by mapping extracted relations such as who tweets to whom in which language, into graph markup formats for analysis in network visualization tools.

DATool Description

- **All Darija**: Fully highlighted tweet
- **Partial Darija**: Red check mark
- **Non-Darija**: No highlight

Automatic classifier

Summary statistics including:
- Total # of tweets & conversations
- Tweets per user

Export annotation files in GEOX for use by GNEF, a network visualization tool

GOALS
- Speed up annotation process
- Improve annotation consistency
- Provide a confidence score

Evaluation Using Arabic Script Tweet Data

**Training Set (Arabic script non-tweets)**

<table>
<thead>
<tr>
<th>Darija</th>
<th>Hesspress</th>
<th>Naskh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query results</td>
<td>Commentary</td>
<td>917</td>
</tr>
</tbody>
</table>

**Test Set (Arabic script tweets)**

<table>
<thead>
<tr>
<th>Non-Darija</th>
<th>Darija</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>387</td>
<td>1,013</td>
<td>1,400</td>
</tr>
</tbody>
</table>

Analysis of Annotated Conversations

- Social Network:
  - Nodes: Users
  - Edges: Tweets between users
  - 135 nodes, 234 edges
  - Colored by in-degree

- Information Network:
  - Nodes: Tweets
  - Edges: replies between tweets
  - 995 nodes, 679 edges
  - Red: Darija Romanized
  - Green: Darija, Arabic
  - Blue: non-Darija

Conclusion and Future Work

- DATool supports semi-automated annotation of tweet conversations
- Ongoing effort to expand classifier to include other Arabic script languages (e.g., Urdu, Farsi)
- When to adjust confidence threshold in DAT? What classified results should a native speaker review?
- Would other network diagrams be more useful in predicting code-switching and providing deep linguistic analysis?

<table>
<thead>
<tr>
<th>Source</th>
<th>MSA</th>
<th>Egyptian</th>
<th>Gulf</th>
<th>Levantine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZCB</td>
<td>63,555</td>
<td>18,875</td>
<td>20,730</td>
<td>11,346</td>
<td>114,506</td>
</tr>
</tbody>
</table>

Classifier Results

Without LDA-derived features

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.9</td>
<td>24.1</td>
</tr>
</tbody>
</table>

With LDA-derived features

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>97.2</td>
<td>44.1</td>
</tr>
</tbody>
</table>