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

Stephen Tratz†, **Douglas Briesch**†, **Jamal Laoudi**‡, and **Clare Voss**†
†Army Research Laboratory, Adelphi, MD 20783
‡ArtisTech, Inc., Fairfax, VA 22030
{stephen.c.tratz.civ,douglas.m.briesch.civ,jamal.laoudi.ctr,clare.r.voss.civ}@mail.mil

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 described briefly herein.

The DATOOL supports downstream discourse analysis of tweeted "conversations" by mapping extracted relations such as, *who tweets to whom in which language,* into graph markup formats for analysis in network visualization tools.

1 Overview

For historically unwritten languages, few textual resources exist for developing NLP applications such as machine translation engines. Even when audio resources are available, difficulties arise when converting sound to text (Robinson and Gadelii, 2003). Increasingly, however, with the widespread use of mobile phones, these languages are being written in social media such as Twitter. Not only can these languages be written in multiple scripts, but conversations, and even individual messages, often involve multiple languages. To build useful textual resources for documenting and translating these languages (e.g., bilingual dictionaries), tools are needed to assist in language annotation for this noisy, multiscript, multilingual form of communication.

This paper presents the Dialect Annotation Tool (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 North African Arabic dialect spoken by millions but not taught in schools and lacking in standardardization and linguistic resources. The DATOOL can retrieve the conversation for each tweet on a user's timeline or via Apollo (Le et al., 2011) and display the discourse, enabling annotators to make more informed decisions. It has integrated classifiers for automatically annotating data so a user can either verify or alter the automatically-generated annotations rather than start from scratch. The tool can also export annotated data to GEPHI (Bastian et al., 2009), an open source network visualization tool with many layout algorithms, which will facilitate future "code-switching" research.

2 Tool Description

2.1 Version 1.0

The first version of the tool is depicted in Figure 1. It is capable of loading a collection of tweets and extracting the full conversations they belong to. Each conversation is displayed within its own block in the conversation display table. An annotator can mark multiple tweets as Darija (or other language) by selecting multiple checkboxes in the lefthand side of the table. Also, if a tweet is written in multiple languages, the annotator can annotate the different sections using the *Message* text box below the conversation display table.

The tool also calculates user and collection level summary statistics, which it displays below the main annotation section.

We worked with a Darija-speaking annotator during the tool's development, who provided valuable feedback, helping to shape the overall design of the tool and improve its functionality.

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 135–139,
Sofia, Bulgaria, August 8-9, 2013.

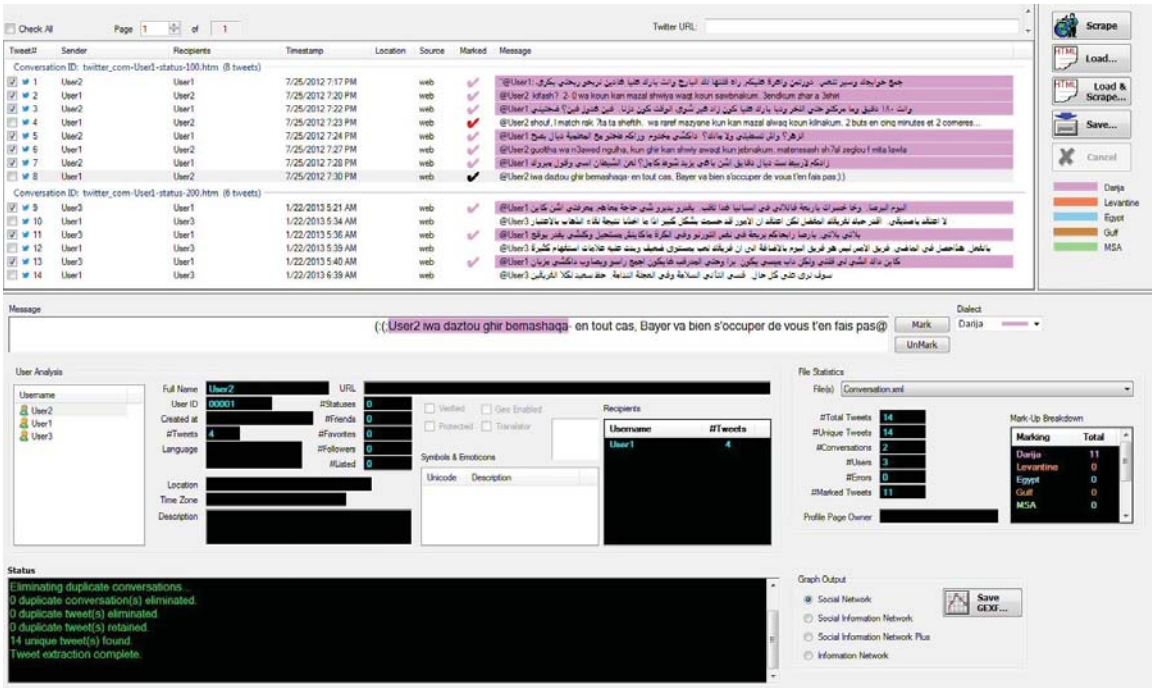

Figure 1: The Dialect Annotation Tool (DATOOL) displaying a possible Twitter conversation.

Data Annotation Using version 1.0, the annotator marked up 3013 tweets from 3 users for the presence of the Darija (approximately 1,000 per user), averaging about 250 tweets per hour. Of the 1,400 tweets with Arabic script, 1,013 contained Darija. This annotated data is used to evaluate the Arabic dialect classifier discussed in Section 3.

2.2 Version 2.0

The second version of the tool contains the additional ability to invoke pre-trained classification models to automatically annotate tweets. The tool displays the classifier's judgment confidence next to each tweet, and the user can set a minimal confidence threshold, below which automatic annotations are hidden. Figure 2 illustrates the new classification functionality.

2.3 XML Output

The DATOOL stores data in an XML-based format that can be reloaded for continuing or revising annotation. It can also export four different views of the data in Graph Exchange XML Format (GEXF), a format that can be read by GEPHI. In the *social network* view, users are represented by nodes, and tweets are represented as directed edges between the nodes. The *information network* view displays tweets as nodes with directed edges between time-ordered tweets within a conversation. In the *social-information network* view, both users and tweets are represented by nodes, and there are directed edges both from tweet senders to their tweets and from tweets to recipients. The *social-information network plus* view provides all the information of both the social network and the information network.

3 Classifier

For the second version of the DATOOL, we integrated an Arabic dialect classifier capable of distinguishing among Darija, Egyptian, Gulf, Levantine and MSA with the goal of improving the speed and consistency of the annotation process.

Though language classification is sometimes viewed as a solved problem (McNamee, 2005), with some experiments achieving over 99% accuracy (Cavnar and Trenkle, 1994), it is significantly more difficult when distinguishing closely-related languages or short texts (Vatanen et al., 2010; da Silva and Lopes, 2006). The only language classification work for distinguishing between these closely-related Arabic dialects that we are aware of was performed by Zaidan and Callison-Burch (2013). They collected web commentary data written in MSA, Egyptian, Levantine, and Gulf and performed dialect identification experiments, their strongest classifier achiev-

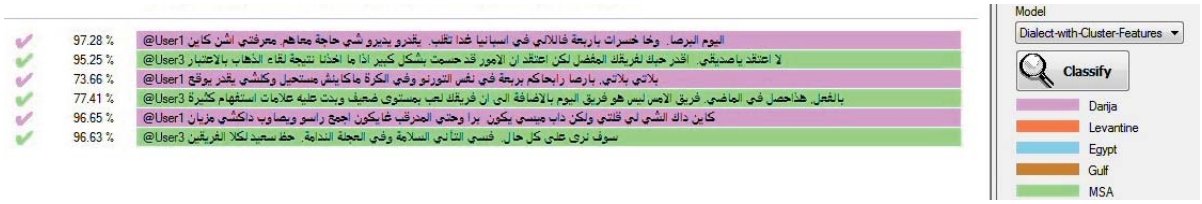

Figure 2: Screenshot showcasing the automatic classification output, including confidence values.

ing 81.0% accuracy.

3.1 Training Data

Since Zaidan and Callison-Burch's dataset includes no Darija, we collected Darija examples from the following sources to augment their dataset: Moroccan jokes from `noktazwina.com`, web pages collected using Darija-specific query terms with a popular search engine, and 37,538 Arabic script commentary entries from `hespress.com` (a Moroccan news website).

Nearly all the joke (N=399) and query term (N=874) data contained Darija. By contrast, the commentary data was mostly MSA. To extract a subset of the commentary entries most likely to contain Darija, we applied an iterative, semi-supervised approach similar to that described by Tratz and Sanfilippo (2007), in which the joke and query term data were treated as initial seeds and, in each iteration, a small portion of commentary data with the highest Darija scores were added to the training set. After having run this process to its completion, we examined 131 examples at intervals of 45 from the resulting ranked list of commentary. The 62nd example was the first of these to have been incorrectly classified as containing Darija. We thus elected to assume all examples up to the 61st of the 131 contain Darija, for a total of 2,745 examples (61*45=2,745). As an additional check, we examined two more commentary entries from each of the 61 blocks, finding that 118 of 122 contain Darija.

3.2 Initial Classifier

The integrated dialect classifier is a Maximum Entropy model (Berger et al., 1996) that we train using the LIBLINEAR (Fan et al., 2008) toolkit. In preprocessing, Arabic diacritics are removed, all non-alphabetic and non-Arabic script characters are converted to whitespace, and sequences of any repeating character are collapsed to a single character. The following set of feature templates are applied to each of the resulting whitespace-separated tokens:

- The full token
- 'Shape' of the token—all consonants are replaced by the letter *C*, alefs by *A*, and *waws* and *yehs* by *W*
- First character plus the last character (if length ≥ 2)
- Character unigrams, bigrams, and trigrams
- The last character of the token plus the first character of the next token
- Prefixes of length 1, 2, and 3
- Indicators that token starts with *mA* and
  - ends with *$*
  - the next token ends with *$*
  - is length 5 or greater

3.3 LDA Model

As an exploratory effort, we investigated using Latent Direchlet Allocation (LDA) (Blei et al., 2003) as a method of language identification. Unfortunately, using the aforementioned feature templates, LDA produced topics that corresponded poorly with the training data labels. But, after several iterations of feature engineering, the topics began to reflect the dialect distinctions. Our final LDA model feature templates are listed below.

- The full token
- Indicators that the token contains
  - *theh*; *thal*; *zah*; *theh*, *thal*, **or** *zah*
- Indicators the token is of length 5+ and starts with
  - *hah* plus *yeh*, *teh*, *noon*, or *alef*
  - *seen* plus *yeh*, *teh*, *noon*, or *alef*
  - *beh* plus *yeh*, *teh*, *noon*, or *alef*
  - *ghain* plus *yeh*, *teh*, or *noon*
  - or *kaf* plus *yeh*, *teh*, or *noon*
- Indicators that token starts with *mA* and
  - ends with *$*
  - the next token ends with *$*
  - is length 5 or greater

The following features produced using the LDA model for each document are given to the Maximum Entropy classifier: 1) indicator of the most-likely cluster, 2) product of scores for each pair of clusters.

3.4 Classifier Evaluation

We evaluated the versions of the classifier by applying them to the annotated data discussed in

Section 2.1. The initial classifier without the LDA-derived features achieved 96.9% precision and 24.1% recall. The version with LDA-derived features achieved 97.2% precision and 44.1% recall, a substantial improvement. Upon review, we concluded that most cases where the classifier "incorrectly" selected the Darija label were due to errors in the gold standard.

4 Analysis of Annotated Conversations

Visualization of Darija in Conversations

The DATOOL may recover the conversation in which a tweet occurs, providing the annotator with the tweet's full, potentially-multilingual context. To visualize the distribution of Darija[1] by script in ≈1K tweets from each user's conversations, the DATOOL transforms and exports annotated data into a GEXF information network (cf. Figure 3), which can be displayed in GEPHI.[2] Currently, GEPHI displays at most one edge between any two nodes—GEPHI automatically augments the edge's weight for each additional copy of the edge.

The Darija in this user's conversations, unlike our two other users, is predominantly Romanized. With more data, we plan to assess the impact of one user's script and language choice on others.

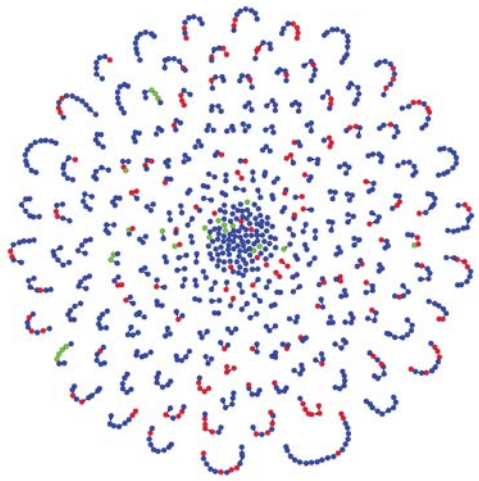

Figure 3: Information network visualization. *Red*—contains Romanized Darija; *green*—contains Arabic-script Darija; *blue*—no Darija.

Code-Switching

The alternation of Darija with non-Darija in the information network (red and green nodes vs. blue nodes) within conversations is consistent with well-known code-switching among Arabic speakers, extending spoken discourse into informal writing (Bentahila and Davies, 1983; Redouane, 2005). Code-switching also appears within our tweet corpus where Romanized Darija frequently alternates with French. Given the prevalence of code-switching within tweets, future work will entail training a Roman-script classifier at the token level.[3] Since our DATOOL already supports token-level as well as multi-token, tweet-internal annotation in the mid-screen *Message* box, our current corpus provides a seed set for this effort.

5 Conclusion and Future Work

The DATOOL now supports semi-automated annotation of tweet conversations for Darija. As we scale the process of building low-resource language corpora, we will document its impact on annotation time when few native speakers are available, a condition also relevant and critical to preserving endangered languages. We have begun extending the classifier to support additional Arabic script languages (e.g., Farsi, Urdu), leveraging resources from others (Bergsma et al., 2012).

Many other open questions remain regarding the annotation process, the visualizations, and the human expert. Which classified examples should the language expert review? When should an annotator adjust the confidence threshold in the DATOOL? For deeper linguistic analysis and code-switching prediction, would seeing participants and tweets, turn by turn, in network diagrams such as Figure 4 help experts understand new patterns emerging in tweet conversations?

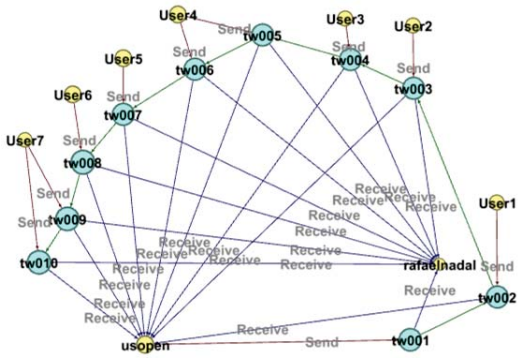

Figure 4: Social-Information Network Plus.

[1] In our initial annotation work, words and tweets in languages other than Darija received no markup.

[2] GEPHI's Force Atlas layout automatically positions subgraphs by size, with larger ones further away from the center.

[3] As described in Section 3, our current classifier works at the tweet level and only on Arabic-script tweets.

Acknowledgments

We would like to thank Tarek Abdelzaher for all his feedback regarding our work and guidance in using Apollo. We would also like to thank our reviewers for their valuable comments and suggestions.

References

Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*.

Abdelali Bentahila and Eirlys E Davies. 1983. The Syntax of Arabic-French Code-Switching. *Lingua*, 59(4):301–330.

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language Identification for Creating Language-Specific Twitter Collections. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pages 65–74.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

William B Cavnar and John M Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Joaquim Ferreira da Silva and Gabriel Pereira Lopes. 2006. Identification of document language is not yet a completely solved problem. In *Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, pages 212–212. IEEE.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Hieu Khac Le, Jeff Pasternack, Hossein Ahmadi, M. Gupta, Y. Sun, Tarek F. Abdelzaher, Jiawei Han, Dan Roth, Boleslaw K. Szymanski, and Sibel Adali. 2011. Apollo: Towards factfinding in participatory sensing. In *IPSN*, pages 129–130.

Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Rabia Redouane. 2005. Linguistic constraints on codeswitching and codemixing of bilingual Moroccan Arabic-French speakers in Canada. In *ISB4: Proceedings of the 4th International Symposium on Bilingualism*, pages 1921–1933.

Clinton Robinson and Karl Gadelii. 2003. Writing Unwritten Languages, A Guide to the Process. *http://portal.unesco.org/education/en/ev.php-URL_ID=28300&URL_DO=DO_TOPIC&URL_SECTION=201.html*, UNESCO, Paris, France. December.

Stephen Tratz and Antonio Sanfilippo. 2007. A High Accuracy Method for Semi-supervised Information Extraction. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 169–172.

Tommi Vatanen, Jaakko J Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC'10*.

Omar Zaidan and Chris Callison-Burch. 2013. Arabic dialect identification. *Computational Linguistics (To Appear)*.