NLP for Arabic

20 March 2017

Arabic Computational Linguistics

Arabic is a Semitic language spoken by “as many as 422 millions of people” [1] in the world with many dialects. Formation of Modern Standard Arabic (MSA), differing from classical Arabic, dates back to 19th century and is said to be ongoing [2].

The language has its own alphabet, is written from right to left and is case-invariant (no lower or upper case letters). It is morphologically complex and as for its grammar we couldn’t find many special remarks about it but due to its morphological structure, the syntactic information is burdened also on the words, making most of the language processing tasks difficult.

Arabic is a relatively low-resource language, especially in terms of corpora [3, 4], in the scope of NLP. Most of the tools and corpora are based on MSA and the studies on dialectical Arabic seems to have risen recently [5], mostly by the growth of social media. Indeed, that the NLP tools are trained on MSA data has been indicated to be a challenge for the possibly pervasive use of those tools as the Arabic world in reality speaks many other dialects other than MSA [6, 7]. It is proposed that language / dialect identification / disambiguation can be applied prior to the final prediction application [9, 10].

There are many studies for sentiment classification in Arabic texts [4, 6, 11, 12]. The major bottleneck is indicated to be the dialect differences in the texts. Other than this, performance values are promising and exceeds those of English and Turkish for example.

A number of spam classification studies can be found [8] while the datasets seem not available.

[1] https://en.wikipedia.org/wiki/Arabic
[2] Farghaly, Ali. "The Arabic language, Arabic linguistics and Arabic computational linguistics." *Arabic Computational Linguistics* (2010): 43-81.
[3] Zaghouani, Wajdi. "Critical survey of the freely available Arabic corpora." *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC*. 2014.
[4] Al-Moslmi, Tareq, et al. "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis." *Journal of Information Science* (2017): 0165551516683908.
[5] Shoufan, Abdulhadi, and Sumaya Al-Ameri. "Natural language processing for dialectical Arabic: A Survey." *ANLP Workshop 2015*. 2015.
[6] Abdul-Mageed, Muhammad, Mona Diab, and Sandra Kübler. "SAMAR: Subjectivity and sentiment analysis for Arabic social media." *Computer Speech & Language* 28.1 (2014): 20-37.
[7] Abdulla, Nawaf A., et al. "Arabic sentiment analysis: Lexicon-based and corpus-based." *Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference on*. IEEE, 2013.
[8] Alsmadi, Izzat, and Ikdam Alhami. "Clustering and classification of email contents." *Journal of King Saud University-Computer and Information Sciences* 27.1 (2015): 46-57.
[9] Dias Cardoso, Pedro Miguel, and Anindya Roy. "Language Identification for Social Media: Short Messages and Transliteration." *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.
[10] Arabic Language Disambiguation for NLP Applications, http://innovation.columbia.edu/technologies/cu14012_arabic-language-disambiguation-for-natural-language-processing-applications
[11] Korayem, Mohammed, David Crandall, and Muhammad Abdul-Mageed. "Subjectivity and sentiment analysis of arabic: A survey." *International Conference on Advanced Machine Learning Technologies and Applications*. Springer Berlin Heidelberg, 2012.
[12] Rushdi-Saleh, Mohammed, et al. "OCA: Opinion corpus for Arabic." *Journal of the American Society for Information Science and Technology* 62.10 (2011): 2045-2054.

Resources for Arabic

Corpora

- Sentiment
    - Multi-domain arabic sentiment corpus (Al-Moslmi et al., 2017) https://github.com/almoslmi/masc/blob/master/MASC%20Corpus.zip
    - Arabic tweets (pos, neg, neutral) http://www.mohamedaly.info/datasets/astd
- Newspaper
    - http://zeus.cs.pacificu.edu/shereen/research.htm

        - Available for research only
- Topic identification corpus
    - https://sites.google.com/site/mouradabbas9/corpora
- Spam
    - Texts not available: https://sites.google.com/site/heiderawahsheh/home/web-spam-2011-datasets/uk-2011-web-spam-dataset
- Misc
    - King Abdulaziz City for Science and Technology (KACST) Arabic corpus: http://www.kacstac.org.sa/
        - Site not available
    - http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm
    - http://users.dsic.upv.es/~ybenajiba/downloads.html
    - http://www.medar.info/BLARK/unannotated_corpora.php

Tools

- Toolkits
    - Farasa: Segmenter, POS tagger, diacritizer, NER, dependency & constituency parser
        - http://qatsdemo.cloudapp.net/farasa/
        - Java-based
    - MADAMIRA: Tokenization, lemmatization, NER, phrase extraction
        - http://camel.abudhabi.nyu.edu/madamira/?locale=en
        - Non-commercial licence
    - Comparable text miner: morphological analysis, document pre/processing, n-gram, tfidf weighting, POS tagging, similarity
        - https://github.com/motazsaad/comparable-text-miner
        - Includes english

- tokenization
    - Nltk - wordpunct_tokenize
        - http://www.nltk.org/_modules/nltk/tokenize/regexp.html#WordPunctTokenizer
    -

- Stemming
    - Nltk ISRI stemmer:
        - http://www.nltk.org/api/nltk.stem.html#nltk.stem.api.StemmerI
        - Available
    - A novel root based Arabic stemmer, 2015
        - Couldn't find the files
    - Kohja's stemmer

        - http://zeus.cs.pacificu.edu/shereen/research.htm
        - Java-based, codes are closed, there is only a gui.
            - Jar version: https://github.com/motazsaad/khoja-stemmer-command-line
    - Lucene
        - https://lucene.apache.org/core/3_6_0/api/contrib-analyzers/org/apache/lucene/analysis/ar/ArabicStemmer.html
        - Java-based
    - Light stemmer - Tashaphyne
        - https://pypi.python.org/pypi/Tashaphyne/
    - Light10 stemmer
        - https://github.com/motazsaad/arabic-light-stemmer

- Morphological analysis
    - MadaMira: https://camel.abudhabi.nyu.edu/madamira/
        - Only online; requires contact for commercial use

    - Standard Arabic Morphological Analysis: https://catalog.ldc.upenn.edu/LDC2010L01
        - Licence problem
        - Couldn't find the files for the tool
    - SAFAR, morphosyntactic tool
        - http://arabic.emi.ac.ma/safar/?q=applications
        - java-based
        - available?

- Spellchecking
    - Yaraspell: https://github.com/linuxscout/yaraspell/blob/master/yaraspell/spelltools.py
    - Aspell: https://github.com/WojciechMula/aspell-python
    - Hunspell: https://github.com/hunspell/hunspell

- NER
    - Polyglot
    - NERaR: https://github.com/SouhirG/NERAr
        - BSD licence
    - AQMAR Arabic tagger:
        - http://www.cs.cmu.edu/~ark/ArabicNER/
        - GPL but has stricter dependencies
        - Run jar file with ?

```
def run_command(command):
    p = subprocess.Popen(command,
                         stdout=subprocess.PIPE,
```

```
                                    stderr=subprocess.STDOUT)
            return iter(p.stdout.readline, b'')
```

- POS tagger
    - Syntaxnet: https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md
    - Turbo parser: http://www.cs.cmu.edu/~afm/TurboParser/README
        - Has a python wrapper
        - LGPL
    - A tagger implemented by the new CRF model: https://wapiti.limsi.fr/
        - BSD licence
    - Stanford: http://nlp.stanford.edu/software/tagger.html
        - Licence problem
    - http://alt.qcri.org/tools/ArabicPOSTaggerLib/
        - Java based
        - LGPL; says it requires contact for commercial use

- Many links to a number of tools for various languages
    - http://multital.inalco.fr/

- WordNet
    - http://arabic.emi.ac.ma/ibtikarat/?q=Resources
        - Availability unclear.

- Keyphrase extraction
    - http://www.claes.sci.eg/coe_wm/kpminer

- Tasks
    - Text normalization?
    - Spam?

People / Groups

- http://www1.ccls.columbia.edu/~mdiab/
- http://www1.ccls.columbia.edu/~ybenajiba/downloads.html
- http://alt.qcri.org/tools/
- http://www.comp.leeds.ac.uk/arabic/
- https://nyuad.nyu.edu/en/research/faculty-research/camel-lab.html
- http://www.ifao.egnet.net/axes-2012/ecritures-langues-corpus/2012-tala/
- http://www1.ccls.columbia.edu/~cadim/
- https://sites.google.com/site/mouradabbas9/
- https://sites.google.com/site/anlprg/

Classification experiments

- Sentiment analysis
    - We studied a baseline model for Arabic sentiment analysis at document-level using a published, human-annotated dataset.

    - Dataset:
        - We used the dataset collected by [1].
        - It has customer reviews collected online from three domains: political, software and mix of 15 other domains like culture, sports.
        - In total there are 5408 positive and 3453 negative reviews; annotated by two people.
        - The texts in this dataset are from various dialects.
    - Features & tools
        - Preprocessing
            - Tokenization: nltk's wordpunct_tokenizer
            - Stemming: nltk's ISRI stemmer
            - Numbers removed
            - Punctuation removed
            - Deasciifying: not applicable for Arabic
            - Downcasing: not applicable for Arabic
        - Features
            - Tfidf weights of word unigrams and bigrams; tfidf weights of character bigrams
    - Classifier
        - We classified with both Naive Bayes and SVM (using SGD) separately.

    - Results

Classification with 5-fold cross-validation on each domain and their .

Dataset (domain)	Number of texts	Accuracy (NB)	Accuracy (SVM)
Software	1204	0.87	0.89
Political	1204	0.83	0.84
Mix	6733	0.88	0.93

All-3-sets	9141	0.87	0.90

- In [2], it is indicated that without stemming, the classification tasks for Arabic where the features are tokens might perform better. We did some tests in this regard but didn't see much improvement.

- Email / message
    - We couldn't find a proper dataset yet.

Classification ToDo

- Sentiment
    - Use lexicons
- News classification

Resources

[1] Al-Moslmi, Tareq, et al. "Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis." *Journal of Information Science* (2017): 0165551516683908.
[2] Rushdi-Saleh, Mohammed, et al. "OCA: Opinion corpus for Arabic." *Journal of the American Society for Information Science and Technology* 62.10 (2011): 2045-2054.

Remarks

- The dialect of the corpus is said to be very important and so the dialect of the target country
    - We need more information about the daily Arabic / Arabic in business; is it in MSA or local?

ToDo

- Sentiment analysis
    - with texts in MSA
        - Train on MSA, test on dialect and vice versa
    - Datasets from social media
    - Lexicon-based, sentence-level, POS tag features
- Finding more datasets
- Deciding on the NLP tools like POS tagger