Keynote paper

Language Technologies for Social Media

Wajdi Zaghouani
Carnegie Mellon University Qatar
Education City, Doha, Qatar
wajdiz@cmu.edu

Summary

We are witnessing an increased interest from stakeholders to collect and analyze in real time the large-volume of information from social media streams using all kinds of applications ranging from information extraction tools to social media analytics and decision support systems. Social media text is generally noisy, short and linguistically rich as witnessed by the high-frequency rate of code-switching and colloquial expressions used. In this paper, I present an overview of the language technologies within the context of social media, and will discuss the data collection and annotation of social media content. Afterwards, several text processing tools and techniques used before building social media applications will be presented. Finally, some social media applications and their evaluation benchmarks are explored.

Key words: Language Technologies, Social Media, Text Processing, Corpus Annotation

Introduction

On-line social networking has revolutionized the way we communicate. Recent research on social media has revealed the impact of social media on the lives of millions of people. Language technologies could help process social media data using the most recent techniques and algorithms to reveal insightful information from the multilingual big data available online (Pouliquen et al. 2006; Zaghouani 2014). We present an overview of the Natural language processing within the context of Social media. First, we will discuss the data collection and annotation of social media content. Afterwards, we will explain the main challenges faced during the text pre-processing of social media text. Finally, we will explore some tools and applications related to social media and their evaluation benchmarks.

Social Media Data Collection and Annotation

In order to build natural language processing tools and systems, training data is needed (Jeblee et al. 2014; Zaghouani et al. 2015). Social media popularity is increasing and a large amount of public user-generated content is becoming available for collection. However, the collection and the annotation of social

media textual data need to be carefully considered for each task before starting the collection effort. Moreover, the data should also be annotated in a consistent way (Maamouri et al. 2010; Zaghouani et al. 2014; Zaghouani et al. 2015).
Social media data collection depends on the planned task and its applications. For instance, social media textual data could be collected in multiple forms such as image descriptions, videos, posts and metadata as explained in (Ford and Voegtlin, 2003). Furthermore, social media data is often full of spam that should be detected and removed from the dataset.
In order to collect social media data, there exist application programming interface (API) used to integrate with other applications (Obeid et al. 2013). However, some restrictions are possible, for instance, the Twitter API has a limitation per user, per the number of the Tweets to be collected and per the application. This will lead to a limited number of requests. Those interested in getting a larger volume of data may opt for paid access.
The annotation of social media content is a challenging task and clear guidelines should be provided to the human annotators (Zaghouani et al. 2016d). In general, a minimum of two annotators is needed for a given task and the guidelines should clearly explain what and how to annotate (Zaghouani et al. 2016e). In order to access the quality of the annotation, inter-annotator measures are frequently performed to check the agreement rate between the annotators (Zaghouani and Dukes 2014). In case of disagreements, the issue is resolved by taking the majority vote of the annotators and for this reason, it is advised to have an odd number of annotators (Bouamor et al. 2015; Zaghouani et al. 2012). To improve the agreement score, the annotators are encouraged to discuss any disagreement until they reach an agreement. The inter-annotator agreement is measured using the kappa statistical measure used to compensate the agreement obtained for possible agreements due to chance (Artstein and Poesio, 2008, Carletta, 1996).
The annotation tasks can also be performed in a semi-automatic way through intelligent interfaces between the annotations and the users as in the case of GATE (General Architecture for Text Engineering) and TwitIE, a related social media tool used for corpus annotation (Bontcheva et al., 2013).

Social Media Text Processing Tools and Techniques

Social media text is full of useful information, however, it is usually informal and written in a naturally occurring way such as the abbreviations in SMS phone messages. The occurrence of non-standard words and misspelled text poses a big challenge for natural language processing (Pouliquen et al. 2005). In order to build language technologies applications for social media, the collected text should go through various normalization steps (Zaghouani et al. 2016b). Text normalization is especially needed to reduce the linguistic noise from the data (Diab et al. 2018; Zaghouani and Awad 2016). Furthermore, the normalization will reduce the linguistic ambiguity in a language such as Arabic (Haw-

wari et al. 2013; Draffan et al. 2015; Zaghouani et al. 2016c). During the normalization process, the orthographic errors are identified and later on they could be corrected using a dictionary of correctly spelled terms. The dictionary generally allows the detection of out-of-vocabulary entries and unknown words.
Natural language processing tools are essential in language technologies projects especially those involving data annotation (Maamouri et al. 2012; Obeid et al. 2016). We identified several tools frequently used for social media text processing and tools specifically developed for social media.

- **The Stanford CoreNLP:** this is a suite of Natural Language Processing tools for the English language. It supports tokenization, parsing, part-of-speech tagging, and named entity recognition. The Stanford POS tagger was trained for social media text by Derczynski et al. (2013a)
- **Open NLP:** this suite supports various functions from tokenization to sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. The OpenNLP chunker by Ritter et al. (2011) was trained specifically also for social media text.
- **FreeLing:** a set of tools for a variety of languages including English. It can do the text tokenization, sentence splitting, morphological analysis, phonetic encoding, named entity recognition, POS tagging, parsing and co-reference resolution.
- **NLTK**: this is a well-known suite of text processing libraries written in Python for classification, tokenization, stemming, POS tagging, parsing, and semantic reasoning task.
- **GATE**: another well-known toolkit that includes various language processing components such as parsers, morphological analyzer, Part-of-speech tagging. It also contains information retrieval tools, information extraction components for various languages among others. Gate has been adapted to social media text processing through the TwitIE module (Derczynski et al., 2013b). This module supports the tokenization of Twitter texts and also the POS tagging and the named entities recognition.
- **NLPTools:** this is a Natural Language Processing library dedicated to text classification, tokenizing, stemming and clustering.
- **TweetNLP:** this part-of-speech tagger was developed at Carnegie Mellon University and was built especially for social media texts (Owoputi et al., 2013). It was created with manually labeled POS annotated tweets. A dedicated Web-based annotation tool was used in this project
- **TweeboParser**: A dependency parser was built using the Twitter annotated Treebank for 929 tweets (Kong et al., 2014).
- **The University of Washington (UW)** Twitter NLP Tools: this is a suite of tools created by (Ritter et al., 2011) and includes a POS tagger and an annotated Twitter data.

Since social media messages are available in multiple languages and in some cases, there is a situation of code-switching, for example, some users may write in Arabic and write part of the text in English. In order to detect the language of social media text, several language detection systems were built.
In order to build these language identification tools for social media, existing tools need to be re-trained and the performances are generally lower due to nature of the social media messages. For instance, language identification systems can achieve around 98% of precision in detecting languages while it will decrease to 90% for data for example as explained in Derczynski et al. (2013a).
Lui and Baldwin (2014) tested several language identification systems on Twitter data and obtained an F-score of 89% with the best system. Twitter dataset became the standard testing and training data used for this kind of tasks. Testing is usually done using existing tools after various text normalization and cleaning steps such as removing hashtags, emoticons, mentions, re-tweets etc..
For language identification task, the methods used relied mostly on the text of the message, but in some cases such as in Carter et al. (2013), they used metadata information, a unique approach in social media. They found that several features can help in identifying the language such as the language profile of the user, the hyperlink content, the language profile of the other users mentioned in the given post, the language of the original post and the language profile of the given tag. They tested their method and it improved by 5% over the baseline.
We identified the following language identification tools:

- **LangDetect**: this is a Bayes classifier and it is based on character n-grams without feature selection and a set of normalization heuristics.
- **Whatlang:** this tool is based on a vector-space model with per-feature weighting over character n-grams (Brown, 2013).
- **Langid.py**: this tool is adapted for more than 90 languages and uses a feature set selection process from various sources (Lui and Baldwin, 2012).
- **LDIG**: this is a Java language identification tool done specifically for Twitter messages. It was trained on 47 languages. It uses a document representation based on data structures.

In some cases, social media posts are written in a dialectal variety and dedicated dialectal identification tools are required in this case. We cite the case of the various Arabic varieties used in social media in 22 Arabic countries. We noticed that dialectal Arabic is usually mixed with standard Arabic in social media messages. In recent years, more attention was given to building applications for Arabic dialectal identification (Zaghouani et al. 2016a).
This task attempts to find dialect variety used in a set of texts that use the same character set in a known language and since dialects within the same language are sometimes very similar, this task is more difficult than language identification. the various machine learning techniques and methods used for language

identification were adapted for dialect identification as well. Once a dialect is identified, it will be mapped to standard Arabic for further processing using the MSA tools as there is a lack of dialect dedicated NLP tools.
We located several projects related to dialectal Arabic, we cite in particular the efforts of (Habash, 2010) and Diab et al. (2010) within the context of the CO-LABA project, a large-scale project to create resources and processing tools for Dialectal Arabic blogs. The project focused mostly on four Arabic dialects: Moroccan, Iraqi, Egyptian and Levantine.

Social Media Language Technologies Application

In this section, we present a selection of some social media related applications based on language technologies. These applications based on social analytic could give useful insights on social media user behavior to small businesses, industry, financial institutions among other institutions.

Health applications

In healthcare, many patients tend to write information about their health and possible treatments and the side effects of medication. They also share their experiences with other social media users. all this data can be useful for health care professionals, for example collecting data about depression could be useful in detecting possible mental health issues. Ali et al. (2013) built a collection of texts from on-line medical groups related to hearing devices and sorted them into positive, neutral or negative. This sentiment annotation is useful for example when we are interested in filtering only messages with a specific opinion. When dealing with health-related data, we need to take into consideration the user privacy and a de-identification process should be performed to remove sensitive personal information from this data.

Financial Applications

In the financial domain, social media analysis can be useful for example in studying the relation between the economic indicators and the financial news and the role of rumors in the stock exchange market fluctuations. Moreover, social media can be used to do surveys and studies and we can cite the example of Twitter data the revealed the public mood of a given population for market research. Sul et al. (2014) collected data from Twitter messages related to companies in the S&P 500 and they analyzed the cumulative emotional valence by comparing the average daily stock market income. Their results revealed that the cumulative emotional valence (negative or positive) of tweets about a specific company was related to a given company stock income. In another application, Bollen et al. (2010) did some analysis on the content of Tweets on a daily basis using mood tracking applications. They measured the negative versus positive mood using six dimensions (sure, calm, alert, vital, happy and kind).

Disaster Relief Applications
Social media messages can be used to monitor and detect signs of an emergency situation in a timely manner for stakeholders in crisis management. For example, a sudden change in trending topics in social media can be a sign of a possible emergency and should be tracked such as early indications of fire, earthquake or Flooding. Also, social media can be used a tool to send updated information about the evolution of the crisis. Language technologies can play a vital role here by the automatic analysis and monitoring of such messages which could help the government agency to quickly react. We cite the work of Yin et al. (2012) who monitored monitor Twitter streams to detect emergency situations by creating an automatic system to enhance situation awareness.

Security and Defence Applications
The massive amount of user-generated social media messages could be vital for safety and security, but it is hard for Humans to manually scroll through these messages in order to detect possible security threats. For example, Terrorists may post their messages and could use social media to spread their views. Security-related social media applications can be applied to find these patterns of suspicious behavior and investigate the suspects profiles such as the work of Mohay et al., (2003) who built an intrusion detection application.

Media Monitoring Applications
Monitoring the online media could be a helpful application for business intelligible and also for computational journalism as it helps interested parties to quickly detect important information that is difficult to get in a traditional way. For instance, these tools can quickly track millions of articles and broadcast media and report in a summary the most meaningful information. For example, Nagarajan et al. (2009) extracted from Twitter the observations on spatial temporal-thematic analysis to real-world events. They used Twitris, a Semantic Web application. Another related application is TwitInfo which can track events on Twitter and collect and visualize in a concise way the events according to the user preference.

Evaluation

In order to evaluate the performance of social media tools and applications, several standard benchmarks were created (Chiao et al. 2006; Temnikova et al. 2016; Rozovskaya et al. 2015; Mohit et al. 2014; Atwell et al. 2010). In recent years, we witnessed a surge in social media related evaluation campaigns such as the annual SemEval campaign,[1] the annual CLEF labs and workshops[2] and

[1] http://alt.qcri.org/semeval2017/index.php?id=tasks
[2] http://www.clef-initiative.eu/

the various iteration of TREC campaign.[3] In EMNLP 2014,[4] a shared task was organized for code-switching detection in Twitter messages and a standard data set was distributed. The data included messages between two languages for four pairs: Chinese-English, Nepalese-English, Spanish-English and Modern Standard Arabic and Arabic dialects.

Conclusion

In this paper, we presented a general overview of the social media text and the language technologies. We started by describing the social media text collection and annotation process, a necessary initial step in any project related to text processing. Later on, we described the text pre-processing step and the NLP tools that are generally used to prepare the data for and build a variety of useful social media real-world applications.

Acknowledgement

This paper was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author.

References

Ali, Tanveer, Marina Sokolova, Diana Inkpen, and David Schramm. Can i hear you? Opinion learning from medical forums. Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), 2013

Artstein, Ron and Massimo Poesio. Inter-coder agreement for computational linguistics. Computational Linguistics, 34:553–596, 2008.

Atwell, Eric, Kais Dukes, Abdul-Baquee Sharaf, Nizar Habash, Bill Louw, Bayan Abu Shawar, Tony McEnery, Wajdi Zaghouani, Mahmoud El-Haj. 2010.Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence. In Grand Challenges in Computing Research for 2010 and beyond. part of ACM-BCS Visions of Computer Science conference. 13-16 April 2010, Edinburgh University

Bollen, Jonah, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Computing Research Repository (CoRR), abs/1010.3003, 2010.

Bontcheva, Kalina, Leon Derczynski, Adam Funk, Mark Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pages 83–90. INCOMA Ltd. Shoumen, BULGARIA, 2013.

Bouamor, Houda, Wajdi Zaghouani, Mona Diab, Ossama Obeid, Kemal Oflazer, Mahmoud Ghoneim and Abdelati Hawwari. A Pilot Study on Arabic Multi-Genre Corpus Diacritization. 2015. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijin, China, July 2015.

Brown, D. Ralf. Selecting and weighting n-grams to identify 1100 languages. In Ivan Habernal and Vaclav Matousek, editors, Text, Speech, and Dialogue, volume 8082 of Lecture Notes in Computer Science, pages 475–483. Springer, 2013

[3] http://trec.nist.gov/

[4] http://emnlp2014.org

Carletta, Jean. Assessing agreement on classification tasks: The kappa statistic. Computatioanl Linguistics, 22(2):249–254, June 1996

Carter, Simon, Wouter Weerkamp, and Manos Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation, 47(1):195–215, March 2013

Chiao, Yun-Chuang, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, Wajdi Zaghouani. Evaluation of multilingual text alignment systems: the ARCADE II project. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Derczynski, Leon, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7-13 September 2013. ACL, 2013a

Derczynski, Leon, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, pages 21–30, Paris, France, May 2013b. ACM

Diab, Mona, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. Colaba: Arabic dialect annotation and processing. In LREC Workshop on Semitic Language Processing, pages 66–74, 010

Diab, Mona, Aous Mansouri, Martha Palmer, Olga Babko-Malaya,Wajdi Zaghouani, Ann Bies, Mohammed Maamouri. A Pilot Arabic Propbank; LREC 2008, Marrakech, Morocco, May 28-30, 2008.

Habash, Nizar. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1):1–187, 2010

Jeblee, Serena; Houda Bouamor; Wajdi Zaghouani; Kemal Oflazer. CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, October 2014.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics.

Lui, Marco and Timothy Baldwin. Accurate language identification of Twitter messages. In Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), pages 17–25, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Lui, Marco and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations, pages 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics

Maamouri, Mohamed, Ann Bies, Seth Kulick, Wajdi Zaghouani, Dave Graff and Mike Ciul. From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In Proceedings of LREC 2010, Valetta, Malta, May 17-23, 2010.

Maamouri, Mohammed, Wajdi Zaghouani, Violetta Cavalli-Sforza, Dave Graff and Mike Ciul. Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement. In Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2012, Montreal, Canada.

Mohay, George, Alison Anderson, Byron Collie, Olivier de Vel, and Rodney McKemmi. Computer and Intrusion Forensics. Artech House, Boston, 2003

Mohit, Behrang, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, Ossama Obeid. The First QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, October 2014.

Nagarajan, Meenakshi, Karthik Gomadam, Amit P. Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In Web Information Systems Engineering - WISE 2009, 10th International Conference, Poznan, Poland, October 5-7, 2009. Proceedings, pages 539–553, 2009

Obeid, Ossama, Houda Bouamor, Zaghouani, Wajdi, Mahmoud Ghoneim, Abdelati Hawwari, Sawsan Alqahtani, Mona Diab, Kemal Oflazer. MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization. In Proceedings of The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media.

Obeid, Ossama, Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Kemal Oflazer and Nadi Tomeh. A Web-based Annotation Framework For Large- Scale Text Correction. In Proceedings of IJCNLP'2013, Nagoya, Japan.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of Human Language Technologies 2013: The Conference of the North American Chapter of the Association for Computational Linguistics, Atlanta, GA, USA, 9-15 June 2013, pages 380–390. ACL, 2013

Peter, Dominey Ford and Thomas Voegtlin. Learning word meaning and grammatical constructions from narrated video events. In Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non Linguistic Data, 2003.

Pouliquen, Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, FlavioFuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best. Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.

Pouliquen, Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani & Jan Žižka. Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres. Available online at: http://edel.univ-poitiers.fr/corela/document.php ?id=490. ISSN 1638-5748.

Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), EMNLP '11, pages 1524–1534, Edinburgh, Scotland, UK., July 2011.

Rozovskaya, Alla, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, Behrang Mohit. The Second QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijing, China, July 2015.

Sul, Keel Hong, Allan R. Dennis, and Lingyao Yuan. Trading on Twitter: the financial information content of emotion in social media. In System Sciences (HICSS), 2014 47th Hawaii International Conference on, pages 806–815, Jan 2014.

Temnikova, Irina, Zaghouani Wajdi, Stephan Vogel, Nizar Habash. 2016. Applying the Cognitive Machine Translation Evaluation Approach to Arabic. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016).

Yin, Jie, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6):52–59, 2012.

Zaghouani, Wajdi and Dana Awad. Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation. In Proceedings of The 2nd workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media.

Zaghouani, Wajdi. Critical Survey of the Freely Available Arabic Corpora. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi, Nizar Habash, Houda Bouamor, Ossama Obeid, Sawsan Alqahtani, Mona Diab and Kemal Oflazer. Filtering Dialectal Arabic Text in Two Large Scale Annotation

Projects. The 2nd Workshop on Noisy User-generated Text (W-NUT), December 11 2016, Osaka, Japan. 2016a.

Zaghouani, Wajdi, Ahmed Abdelali, Francisco Guzman and Hassan Sajjad. Normalizing Mathematical Expressions to Improve the Translation of Educational Content. In Proceedings of the AMTA 2016 Workshop Semitic Machine Translation (SeMaT) Collocated with EMNLP 2016 Workshops on November 1st, 2016 Austin, Texas, USA. 2016b.

Zaghouani, Wajdi, Abdelati Hawwari, Sawsan Alqahtani, Houda Bouamor, Mahmoud Ghoneim, Mona Diab and Kemal Oflazer. Using Ambiguity Detection to Streamline Linguistic Annotation, In Proceedings of Coling Workshop "Computational Linguistics for Linguistic Complexity" (CL4LC), Osaka Japan. 2016c.

Zaghouani, Wajdi, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, Kemal Oflazer. Guidelines and Framework for a Large Scale Arabic Diacritized Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016). 2016d.

Zaghouani, Wajdi, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, Kemal Oflazer. Building an Arabic machine translation post-edited corpus: Guidelines and annotation. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016). 2016e.

Zaghouani, Wajdi, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider and Kemal Oflazer. Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus. In Proceedings of the 9th Linguistic Annotation Workshop, co-located with NAACL in Denver, Colorado, USA, 2015.

Zaghouani, Wajdi, Taha Zerrouki and Amar Balla. SAHSOH@QALB-2015 Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers' Errors. The Second QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijing, China, July 2015.

Zaghouani, Wajdi, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh and Kemal Oflazer. Large-scale Arabic Error Annotation: Guidelines and Framework. in the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi and Kais Dukes. Can Crowdsourcing be used for Effective Annotation of Arabic? In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi, Abdelati Hawwari and Mona Diab. A Pilot PropBank Annotation for Quranic Arabic. In Proceedings of the first workshop on Computational Linguistics for Literature, NAACL-HLT 2012, Montreal, Canada.