FURTHER INVESTIGATIONS ON DEVELOPING AN ARABIC SENTIMENT LEXICON

Omar Abdullah Batarfi, Mohamed Y. Dahab1 and Muazzam A. Siddiqui

Faculty of Computer Sciences and Information Technology,
King Abdulaziz University, Jeddah, KSA

ABSTRACT

The availability of lexical resources is huge to accelerate and simplify the sentiment analysis in English. In Arabic, there are few resources and these resources are not comprehensive. Most of the current research efforts for constructing Arabic Sentiment Lexicon (ASL) depend on a large number of lexical entities. However, the coverage of all Arabic sentiment expressions can be applied using refined regular expressions rather than a large number of lexical entities. This paper presents an ASL that more comprehensive than the existing lexicons, for covering many expressions with different dialects including Franco-Arabic, and in the same time more compact. Also, this paper shows how to integrate different lexicons and to refine them. To enrich lexical entries with very robust morphological syntactical information, regular expressions, the weight of sentiment polarity and n-gram terms have been augmented to each

KEYWORDS

Arabic Natural Language Processing, Arabic Sentiment Lexicon, Sentiment Analysis, Text Mining.

1. INTRODUCTION

A lexicon is a structure of word list defined in a specific natural language, which is called a vocabulary, along with a range of knowledge about words and how each word is used. The purpose of a lexicon may be general or may be used in a specific domain. An Arabic lexicon may include several thousand common words of Arabic. An entry in a lexicon may be a single word or multi-word expressions such as noun phrases, and other common expressions (' عيد أهلا وسهلا', 'سعيد') . Each word or phrase in a lexicon is stored in a lexical entry. The lexical entry may include several properties such as the properties of spelling or sound, morphological information, grammatical behavior and/or meaning. Sentiment analysis is the activity of identifying the polarity of sentiment for a given text written in a specific natural language concerning a specific domain or subject. The essence of sentiment analysis is extremely increasing in the opinionated microblogs such as blogs, reviews, and discussions [1]. lexical resources facilitate most of the existing methods to perform the sentiment analysis of microblogs. So, the more the refining sentiment lexicon, the more accurate the results of sentiment analysis.

In this paper, an ASL is constructed from different existing Arabic lexicons. each entry in the constructed lexicon is described based on their sentiment orientation as positive or negative and by their weight.

In previous work [2], there were several problems such as many redundant entries, need for a stemmer, ambiguous sentiment weight, absence of comprehensive, lack of expectation of extra letters to express deep feeling (e.g. مبروووووك instead of مبروك which means congratulation) and

DOI: 10.5121/ijnlc.2019.8601

a huge number of entries. Accordingly, it is a crucial step to find an Arabic sentiment lexicon representation, which is compact and fast, that enables matching a single entry with many word forms and at the same time facilitating the maintains of the lexicon because of the compacting the lexicon's entries. This could be done by adding the important morphological components to each entry.

To the best of our knowledge, there are only four research efforts for developing Arabic sentiment lexicons [2], [3] and [4]. Most of these research efforts rely on a large number of lexical entities. However, the coverage of all Arabic sentiment terms can be applied using refined regular expressions rather than a large number of lexical expressions. For example, the word سعيد (happy) is stored in a regular expression (؟(ء)د+سعي؟(ال)و؟) which covers all possible inflections and extra letters as well.

The following section provides a literature review. Section 3 presents an overall architecture of the proposed methodology. Section 4 presents the results of the experiments which are used to validate the method. Finally, section 5 is dedicated to the conclusions and future work.

2. RELATED WORK

There has not been much work that creates Arabic Semi-supervised Learning (SSL) with intensity scores. However, if we consider Arabic SSL with intensity scores as a special case of lexicon learning, at that point a lot of relevant research works be assigned as relevant. A great number of research works have been developed to give a demonstration of how to construct a sentiment lexicon for the purpose of distinguishing the sentiment polarity for many natural languages such as German, Dutch, Chinese, Spanish, and Japanese, and most seriously, English. This section is going to focus on related work that accomplished for Arabic, the focus of this paper, and on English where most efforts have been focused. Besides that this section will cover also work carried out regarding language independence.

The orientation algorithm was developed for predicting an adjective by Hatzivassiloglou and McKeown [5]. A proposal method of determining a document's polarity was developed by Turney and Littman [6]. The proposed method includes issuing queries to a Web search engine. A lexical resource such as WordNet [7] is used in [8] [9] [10] [11]. The aforementioned research works started with an initial small set, which is manually selected, and by following the built-in WordNet relations, they were able to expand the initial small subset.

Also, Kim and Hovy [8] began with an initial small set of words, seed lists. The seed list was classified into verbs and adjectives, and each of which category was also classified into positive and negative. The first class includes 44 verbs with 23 positive and 21 negatives, while the second class includes 34 adjectives with 15 positive and 19 negatives. The seed list was allowed to be expanded iteratively by exploiting the WordNet. Synonyms and antonyms relations were used to expand adjectives and only synonyms were used for expanding verbs. The researchers come up with the following results: 5880 positive adjectives, 6233 negative adjectives, 2840 positive verbs, and 3239 negative verbs.

Based on the classification of word glosses, Esuli and Sebastiani [9] exploited the WordNet to determine the orientation of a term by assuming that the terms with similar orientation tend to have similar glosses. Esuli and Sebastiani [10] have improved [12] their method that was developed previously in [9] by exploiting both the determination of term subjectivity and the term orientation together.

Kamps et al. [11] determined sentiments of adjectives in the WordNet by calculating the relative distance of the term from the two seed words "good" and "bad".

Elhawary and Elfeky [13] have constructed an Arabic lexicon by exploiting the similarity graph. The similarity graph is a type of graph where each word or phrase represents a vertex/node in the graph. A relation/edge between two words/phrases is constructed between them if they have similar polarity or meaning. The weight of the relationship represents the degree of similarity between two words/phrases. The proposed method initially used a subset of suggested words known as a seed list. The seed list contains 1600 words divided into 600 positive, 900 negatives, and 100 neutral.

Arabic lexical resources such as Penn Arabic Treebank [14] and SentiStrength project [15] are used in [16] and [17] respectively. Abdul-Mageed and Korayem [16] created an Arabic SSL manually based on Penn Arabic Treebank. The researchers have extracted all adjectives from all the first four parts of the Penn Arabic Treebank and manually selected those adjectives that they believed are either positive or negative.

El-Halees [17] created an Arabic SSL manually based on two resources, the SentiStrength project and an online dictionary. The researcher translated the English list from the SentiStrength project and then manually filtered it. Common Arabic words were added to the lexicon.

The authors in [18] and [19] have exploited a simple machine translation procedure to the existing English polarity lexicon. Elaranoty et al. [18] created an Arabic SSL contains strong as well as weak subjective clues by manually translating the MPQA lexicon [20].

Abdul-Mageed and Diab [19] use a machine translation procedure to translate available English lexicons including SentiWordNet [21]. The SentiWordNet is a subset of WordNet which concerns only for sentiment polarity lexicon. The SentiWordNet is widely used for the English Language. The authors translated the SentiWordNet [22] into Arabic.

Most of all aforementioned systems do not have morphological components to the lexical entries. This leads to having different entries for different morphemes.

3. BUILDING ARABIC SENTIMENT LEXICON

The main idea of building ASL, in this paper, is to extend, integrate and refine the existing ASL's. The integration among existing ASL's gives the developed ASL the features of large scale and coverage for Modern Standard Arabic (MSA). While the refinement process gives the developed ASL the features of a compressed or compact version of ASL and avoiding the noise of unused entries. The development of ASL contains the following steps:

- Integrating the existing ASL's
- Refining the combined lexicon (phase I)
- Transliteration of Arabic entries
- Assigning regular expression for each lexical entry
- Refining the lexicon (phase II)

3.1. Integrating Lexicons

The aim of this phase is to construct a compact version of the subjectivity lexicon from combining data coming from different lexicons [2] and [23]. The aforementioned lexicons have the following problems:

- These lexicons have different representations and structures
- They have a lack to represent the various dialects
- They have different file representations

- They have different structures
- Because the Arabic language is highly inflected, we may find different entries for the same word
- Missing entries problem come up with lack of using stemmer
- Different judgments for the same word
- [2] has 12566 out of 15118 terms are neutral terms
- The same lexicon may have different judgments for the same term or entry

Because [2] has a score for positive and negative terms, initial scores have been added to [23]. The output of this phase, the combined lexicon, has 25059 lexical entries. Each entry has its own positive and negative score.

3.2. Refining the Combined Lexicon (Phase I)

This phase includes the following steps:

a) Sort terms alphabetically
b) Remove repeated terms with the same stem and different morphological forms. For example, سعيد (happy for male), سعيدة (happy for female). In this case, remove the most inflected term.
c) Simplify compound terms. Most of the compound terms in the selected lexicons are not real compound terms. For example, يا ناصح (Oh mentor) is not a subjective compound expression because ناصح as a unigram has already existed in the lexicon, may found in other text such as: ناس نصحة, عيال نصحة, etc. while متصدر لا تكلمنى is a compound expression because none of its parts in the lexicon. Other difficult examples are جت الحزينة تفرح and يافرحة امك بك (oh how your mother would be delighted of you). The expression has a component in positive sentiment and the other one in negative sentiment, but totally, the expression is negative.

3.3. Transliteration of an Arabic entries

There are two kinds of vowels in the Arabic language: long vowels and short vowels. The long vowels are written as letters while the short vowels are not letters but they are written as punctuation marks. The short vowels are naturally neglected in MSA text. This research presents a tool to transliterate an Arabic lexicon using English letters to show up each long and short vowel. The problem of representing Arabic words using regular expression in generating Arabic words is explained in more detail in [24].

Arabic texts need to be supplied with short vowels, i.e. Arabic text should be augmented with diacritics, in order to facilitate the process of transliteration. an Arabic-English transliteration method in which an Arabic word is divided into conversable units that are partial Arabic character strings with short vowels is converted into a partial English character string. The transliteration process is a one-to-one mapping from Arabic to English and from English to Arabic.

In this work, we are using Buckwalter Arabic transliteration with minor modifications. Table (1) shows a snapshot of the lexicon with many examples of Arabic words and their transliterations.
One of the most important motivations to transliterate the Arabic lexicon is the inability of MSA to recognize patterns because MSA does not have diacritics. Also, many Arabic users write text reviews or posts using what is called Franco-Arabic which is in the form of transliterated Arabic. Therefore, using the transliterations of Arabic will help for covering many expressions with different dialects including Franco-Arabic.

Table 1. A Snapshot of The Lexicon with Examples of Arabic Words and Their Transliterations

Terms			
ID	**Arabic**	**English**	**Transliterated**
454308	دِيَر	abodes	dIYar
454309	بَطّل	abolish	baT~al
454310	بَطِّل	abolish	baT~Il
454311	أزاح	abolish	azaA7
454312	أزاح	abolish	azaA7
454313	أزَح	abolish	aza7
454315	زِيح	abolish	zIY7
454316	زِح	abolish	zI7
454317	ناسِخ	abolishing	naAsI5
454318	إبْطال	abolition	IbTaAl
454320	إزاحَة	abolition	IzaA7aP
454321	أجْهَض	abort	ajHaD

3.4. Assigning Regular Expression for each Lexical Entry

To match all inflected forms of lexical entry as well as adding extra letters, the regular expression has been added to each entry. The tool published in [24] has been modified and used in this task. For example, the regular expression of the term فرحان (which means happy) is:

و؟[كب]؟(ال)؟فرحا+ن(ي+ن|ة)؟

For example, figure 1 shows many forms of the same meaning or entry in the lexicon with different morphs and possible mistakes match the aforementioned regular expression.

3.5. Refining the Lexicon (Phase II)

This phase is devoted to answering the following questions:

1 What is the meaning of weight?
2 What is the real weight of lexical entry?
3 Is it feasible to remove more entries?
4 First, the weight may express how much the term holds the strength of meaning for both positive and negative sentiment, for example, excellent and good, or may express how much the term is used frequently in either positive or negative sentiment. In this research, the latter is used.

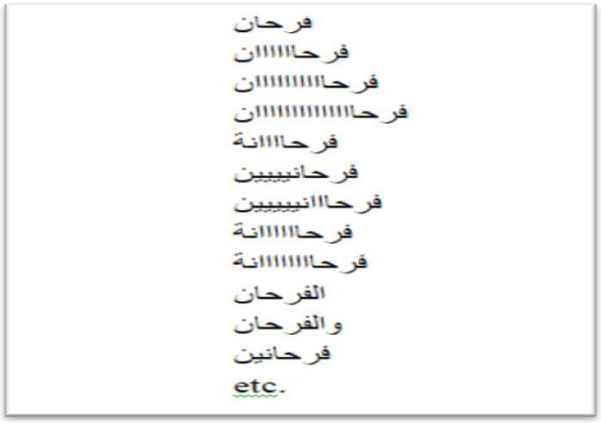

Figure 1. Same entry with many forms in lexicon

Second, the dataset provided in [25] is used to measure how much the term is used frequently in either positive or negative sentiment. Firstly, each term, for the purpose of this task, is represented as the following tuple <term, posFreq, negFreq>. So, a term can be represented as a point in 2D.

Lastly, the ratio of each term frequency has been calculated from total positive documents, 10000 documents, and total negative documents, 10000 documents respectively.

4. EXPERIMENTAL EVALUATION

The term is represented as <term, posProb, negProb>. posProb is the probability of the term to appear in the positive text in the sample. negProb is the probability of the term to appear in negative text in the sample. The average and standard deviation of posProb were 0.021351% and 0.02131% respectively. While the average and standard deviation of negProb were 0.000208%. and 0.02131% respectively.

The experiments show that the constructed lexicon has a limitation for dealing with different Arabic dialects.

5. CONCLUSIONS

This paper presents an ASL that more comprehensive than the existing lexicons, for covering many expressions with different dialects including Franco-Arabic. The ASL is the result of integrating and refinement different existing Arabic sentiment lexicon.

To enrich lexical entries with very robust morphological syntactical information, regular expressions, the weight of sentiment polarity and n-gram terms have been augmented to each entry.

The experiments show that the lexicon is refined, and each term has a great probability to appear in sentiment text especially the positive terms.
We developed an Arabic sentiment lexicon (ASL) which is more comprehensive and refined than existing lexicons. The ASL is the result of integrating and refinement different existing Arabic sentiment lexicon. To enrich lexical entries with very robust morphological syntactical information, regular expressions, the weight of sentiment polarity and n-gram terms have been augmented to each entry. All entries are transliterating to cover Franco-Arabic which may be

written in modern standard Arabic (MSA). By using regular expressions, we cover the misspelling and extra vowel letters that may be added in the middle of words.

For future works, adding more Arabic dialects (such as Egyptian, Tunisian, Moroccan, Levantine, Hejazi , Gulf (Kuwait), Najdi, Gulf (Bahrani), Iraqi, etc.) as well as adding diacritics.

ACKNOWLEDGMENTS

This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH) – King Abdulaziz City for Science and Technology - the Kingdom of Saudi Arabia – award number (12-inf2751-03). The authors also, acknowledge with thanks Science and Technology Unit, King Abdulaziz University for technical support”

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[2] F. Mahyoub, M. Siddiqui and M. Y. Dahab, "Building an Arabic sentiment lexicon using semi-supervised learning," Journal of King Saud University-Computer and Information Sciences, vol. 26, no. 4, pp. 417--424, 2014.

[3] G. Badaro, R. Baly, H. Hajj, N. Habash and W. El-Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining," in Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP), Doha, 2014.

[4] R. Eskander and O. Rambow, "SLSA: A sentiment lexicon for Standard Arabic," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, 2015.

[5] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics, p. 174–181, 1997.

[6] P. D. Turney and M. L. Littman, "Unsupervised learning of semantic orientation from a hundred-billion-word corpus," Technical Report EGB-1094, National Research Council Canada, 2002.

[7] C. Fellbaum, Wordnet, an Electronic Lexical Database, Cambridge, MA: MIT Press, 1998.

[8] S.-M. Kim and E. Hovy, "Determining the Sentiment of Opinions," Proceedings of COLING-04, 20th International Conference on Computational Linguistics, p. 1367–1373, 2004.

[9] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss analysis.," In Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management, p. 617–624, 2005.

[10] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[11] J. Kamps, M. Marx, R. J. Mokken and M. d. Rijke, "using wordnet to measure semantic orientation of adjectives," Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, vol. 4, p. 1115–1118, 2004.

[12] A. Aqel, S. Alwadei and M. Dahab, "Building an Arabic Words Generator," International Journal of Computer Applications, vol. 112, no. 14, pp. 36-41, 2015.

[13] M. Elhawary and M. Elfeky, "Mining Arabic Business Reviews," IEEE International Conference on Data Mining Workshops, p. 1108–1113, 2010.

[14] M. Maamouri, A. Bies, T. Buckwalter and W. Mekki, "The penn arabic treebank: Building a large-scale annotated arabic corpus," in NEMLAR Conference on Arabic Language Resources and Tools, 2004.

[15] M. Thelwall, K. Buckley, G. Paltoglou and D. Cai, "Sentiment Strength Detection in Short Informal Text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, 2010.

[16] M. Abdul-Mageed and M. Korayem, "Automatic identification of subjectivity in morphologically rich languages: the case of Arabic," Proceedings of the 1st workshop on computational approaches to subjectivity and sentiment analysis (WASSA), pp. 2-6, 2010.

[17] A. El-Halees, "Arabic opinion mining using combined classification approach," the international Arab conference on information technology, pp. 10-13, 2011.

[18] M. Elarnaoty, S. AbdelRahman and A. Fahmy, "A Machine Learning Approach For Opinion Holder Extraction Arabic Language," in CoRR, 2012.

[19] M. Abdul-Mageed and M. Diab, "Toward building a large-scale Arabic sentiment lexicon," Proceedings of the 6th International Global WordNet Conference, 2012.

[20] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005.

[21] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available resource for opinion mining," Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), p. 417–422, 2006.

[22] M. Abdul-Mageed, M. Korayem and A. YoussefAgha, ""Yes we can?": Subjectivity Annotation and Tagging for the Health Domain," in Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, Hissar, Bulgaria, 2011.

[23] HARF, "ARABIC LANGUAGE TECHNOLOGY CENTER (ALTEC)," 5 4 2012. [Online]. Available: http://www.altec-center.org/page.php?pg=filesrepository/getRepository.php&main_cat=1&sub_cat=24. [Accessed 1 3 2016].

[24] A. Aqel, S. Alwadei and M. Dahab, "Building an Arabic Words Generator," International Journal of Computer Applications, vol. 112, no. 14, pp. 36-41, 2015.

[25] M. A. Siddiqui, M. Y. Dahab and O. A. Batarfi, "Building A Sentiment Analysis Corpus With Multifaceted Hierarchical Annotation," International Journal of Computational Linguistics (IJCL), vol. 6, no. 2, pp. 11-25, 2015.

[26] C. Fellbaum, M. Alkhalifa, W. J. Black, S. Elkateb, A. Pease, H. Rodr.guez and P. Vossen, "Introducing the Arabic WordNet Project," Proceedings of the 3rd Global Wordnet Conference, 2006.

[27] N. Godbole, M. Srinivasaiah and S. Skiena, "Large-scale sentiment analysis for news and blogs," Proceedings of the International Conference on Weblogs and Social Media ICWSM, 2007.

[28] A. Valitutti, C. Strapparava and O. Stock, "Developing Affective Lexical Resources," PsychNology, vol. 2, no. 1, pp. 61-83 , 2004.

[29] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. U. López and J. M. Perea-Ortega, "OCA: Opinion Corpus for Arabic," Journal of The American Society for Information Science and Technology, vol. 62, no. 10, pp. 2045-2054, 2011.

[30] Y. Yang, "Noise Reduction in a Statistical Approach to Text Categorization," Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 256--263, 1995.

[31] G. Salton, A. Wong and C. S. Yang, "A vector space model for automatic indexing," Commun. ACM , vol. 18, no. 11, pp. 613 - 620, 1975.

[32] M. M. Boudabous, N. C. Kammoun, N. Khedher, L. H. Belguith and F. Sadat, "Arabic WordNet semantic relations enrichment through morpho-lexical patterns," in Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference, Sharjah, 2013.

[33] C. Fellbaum, M. Alkhalifa, W. J. Black, S. Elkateb, A. Pease, H. Rodr.guez and P. Vossen, "Introducing the Arabic WordNet Project," Proceedings of the 3rd Global Wordnet Conference, 2006.

[34] "WordNet 3.0 database statistics," [Online]. Available: https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html#toc. [Accessed 15 4 2013].

AUTHORS

Dr Batarfi received his B.S. degree in Computer Science from King Abdulaziz University, Jeddah, Saudi Arabia in 1989 and his M.S. degree in Artificial Intelligence from George WashingtonUniversity,Washington,D.C.,USAin 1996. He received his Ph.D. from University of Newcastle Upon Tyne, UK in 2008. From 2008 to 2016, he was an Assistant Professor with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. He is currently an Associate Professor of Networking Security at Information Technology Department, Faculty of ComputingandInformation Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include Big Data, Cloud Computing and Information Security.

Dr Dahab is an associate professor at the Department of Computer Science in the Faculty of Computing and Information Technology, King Abdul Aziz University (KAU), Jeddah, Saudi Arabia. He served as the Chairman of the agricultural expert systems development department for 2 years at The Central Laboratory for Agricultural Expert Systems (CLAES), Ministry of Agriculture Egypt. His main research interests include pattern recognition, natural language processing, expert systems, knowledge bases and information retrieval.

Muazzam Ahmed Siddiqui is an associate professor at the Faculty of Computing and Information Technology, King Abdulaziz University. He received his BE in electrical engineering from NED University of Engineering and Technology, Pakistan, and MS in computer science and PhD in modeling and simulation from University of Central Florida. His research interests include sentiment analysis, automatic MCQ generation, entity and relationship extraction and educational data mining.