Sentiment Lexicons for Arabic Social Media

Saif M. Mohammad[1], **Mohammad Salameh**[2], **Svetlana Kiritchenko**[1]
[1]National Research Council Canada, [2]University of Alberta
saif.mohammad@nrc-cnrc.gc.ca, msalameh@ualberta.ca, svetlana.kiritchenko@nrc-cnrc.gc.ca

Abstract

Existing Arabic sentiment lexicons have low coverage—only a few thousand entries. In this paper, we present several large sentiment lexicons that were automatically generated using two different methods: (1) by using distant supervision techniques on Arabic tweets, and (2) by translating English sentiment lexicons into Arabic using a freely available statistical machine translation system. We compare the usefulness of new and old sentiment lexicons in the downstream application of sentence-level sentiment analysis. Our baseline sentiment analysis system uses numerous surface form features. Nonetheless, the system benefits from using additional features drawn from sentiment lexicons. The best result is obtained using the automatically generated Dialectal Hashtag Lexicon and the Arabic translation of the NRC Emotion Lexicon (accuracy of 66.6%). Finally, we describe a qualitative study of the automatic translations of English sentiment lexicons into Arabic, which shows that about 88% of the automatically translated entries are valid for Arabic as well. Close to 10% of the invalid entries are the result of gross mistranslation, close to 40% are due to translation into a related word, and about 50% are due to differences in how the word is used in Arabic.

Keywords: Arabic sentiment lexicons, sentiment analysis, social media, Arabic texts, valence, opinion, translation

1. Introduction

Sentiment lexicons are lists of positive and negative words, optionally with a score indicating the degree of polarity. Sentiment analysis systems, including the best performing ones such as the NRC-Canada system (Mohammad et al., 2013; Kiritchenko et al., 2014a; Zhu et al., 2014), use sentiment lexicons to obtain significant improvements (Wilson et al., 2013; Pontiki et al., 2014; Rosenthal et al., 2015; Mohammad et al., 2016a). However, much of the past work has focused on English texts and English sentiment lexicons.

Arabic sentiment analysis has benefited from recent work (Farra et al., 2010; Abdul-Mageed et al., 2011; Badaro et al., 2014; Refaee and Rieser, 2014), but a resource that is particularly lacking is a large sentiment lexicon. Existing lexicons contain only a few hundred to a few thousand entries. Further, most do not indicate the degree of association between a word and positive (or negative) sentiment. Kiritchenko et al. (2016) created a manually annotated Arabic sentiment lexicon with real-valued sentiment scores; however, that lexicon too has only around one thousand entries.

In this paper, we describe how we created Arabic sentiment lexicons with tens of thousands of entries. These include new automatically generated ones as well as translations of existing English lexicons. We use the lexicons (old and new) for sentiment classification of Arabic social media posts. Our baseline system uses numerous surface form features. Nonetheless, the system benefits from using additional features drawn from sentiment lexicons. We also show the extent to which various sentiment lexicons are effective in sentiment analysis. The best result was obtained using both the Dialectal Hashtag Lexicon and the translated NRC Emotion Lexicon (an accuracy of 66.6%).

Finally, we present a study that qualitatively examines the automatically generated Arabic translations of entries in an English sentiment lexicon. A native speaker of Arabic determined whether the automatic translations were appropriate. An appropriate entry is an Arabic translation that has the same sentiment association as its English source word. Translated entries that were deemed incorrect were further classified into coarse error categories. The study showed that about 88% of the automatically translated entries are valid for Arabic as well. Close to 10% of the errors were caused by gross mistranslations, close to 40% by translations into a related word, and about 50% by differences in how the word is used in Arabic. For further analysis of how translation of words and sentences alters their sentiment, we refer the reader to Mohammad et al. (2016b) and Salameh et al. (2015). All of the lexicons we created are made freely available.[1]

2. Generating Arabic Sentiment Lexicons

We created Arabic sentiment lexicons automatically using two different methods: (1) by using distant supervision techniques on Arabic tweets, and (2) by translating English sentiment lexicons into Arabic using Google Translate—a freely available statistical machine translation system.[2] Even though Google Translate is a phrase-based statistical machine translation system that is primarily designed to translate sentences, it can also provide one-word translations. These translations are often the word representing the predominant sense of the word in the source language. Table 1 lists the number of entries in some of the existing Arabic sentiment lexicons. Table 2 lists the number of entries in each of the lexicons we created. Note that the manually created Kiritchenko et al. (2016) lexicon, and the automatically generated Arabic lexicons (Table 2 - a.i., a.ii., a.iii.) have real-valued sentiment association scores. For the purposes of these tables, terms with scores less than 0 are considered negative and those with scores greater than or equal to 0 are considered positive.

2.1. New Arabic Sentiment Lexicons

The emoticons and hashtag words in a tweet can often act as sentiment labels for the rest of the tweet. We use this idea, commonly referred to as *distant supervision* (Go et al., 2009), to generate three Arabic sentiment lexicons:

[1]http://www.saifmohammad.com/WebPages/ArabicSA.html
[2]Google Translate: https://translate.google.com

- Arabic Emoticon Lexicon: We collected close to one million Arabic tweets that had emoticons (":)" or ":("). For the purposes of generating a sentiment lexicon, ":)" was considered a positive label (*pos*) and ":(" was considered a negative label (*neg*). For each word w, that occurred at least five times in these tweets, a sentiment score was calculated using the formula shown below (proposed earlier in Mohammad et al. (2013) and Kiritchenko et al. (2014b)):

  $$SentimentScore(w) = PMI(w, pos) - PMI(w, neg) \quad (1)$$

  where PMI stands for Pointwise Mutual Information. We refer to the resulting entries as the *Arabic Emoticon Lexicon*.

- Arabic Hashtag Lexicon: The NRC-Canada system used 77 positive and negative seed words to generate the English NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013; Kiritchenko et al., 2014b). We translated these English seeds into Arabic using Google Translate. Among the translations provided, we chose words that were less ambiguous and tended to have strong sentiment in Arabic texts.

  We polled the Twitter API to collect tweets that included these seed words as hashtags. For the purposes of generating a sentiment lexicon, a positive seed hashtag was considered a positive label (*pos*) and a negative seed hashtag was considered a negative label (*neg*). For each word w that occurred at least five times in these tweets, we calculated a sentiment score using Equation 1. We will refer to this lexicon as the *Arabic Hashtag Lexicon*.

- Arabic Hashtag Lexicon (Dialectal): Refaee and Rieser (2014) manually created a small sentiment lexicon of 483 dialectal Arabic sentiment words from tweets. We used these words as seeds to collect tweets that contain them, and generated a PMI-based sentiment lexicon just as described above. We refer to this lexicon as the *Dialectal Arabic Hashtag Lexicon* or *Arabic Hashtag Lexicon (dialectal)*.

In Section 3, we show how we used these lexicons for sentiment analysis.

2.2. Generating Arabic Translations of English Sentiment Lexicons

We used Google Translate to translate into Arabic the words in each of the following English sentiment lexicons: AFINN (Nielsen, 2011), Bing Liu Lexicon (Hu and Liu, 2004), MPQA Subjectivity Lexicon (Wilson et al., 2005), NRC Emotion Lexicon (Mohammad and Turney, 2010; Mohammad and Turney, 2013), NRC Emoticon Lexicon aka Sentiment140 Lexicon (Mohammad et al., 2013; Kiritchenko et al., 2014b), and NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013; Kiritchenko et al., 2014b). Note that Google Translate was unable to translate some words in these lexicons. Table 2 gives the number of words translated (total) as well as a break down by sentiment category (positive, negative, and neutral). In Section 4, we present a study that manually examines a subset of the automatic translations.

3. Arabic Sentiment Analysis

To determine the usefulness of the Arabic sentiment lexicons, we apply them in a sentence-level sentiment analysis system. We build an Arabic sentence-level sentiment analysis system by reconstructing the NRC-Canada English system (Mohammad et al., 2013; Kiritchenko et al., 2014b) to deal with Arabic text. A linear-kernel Support Vector Machine (Chang and Lin, 2011) classifier is trained on the available training data. The classifier leverages a variety of surface-form and sentiment lexicon features. The surface form features include the presence/absence of word and character ngrams, all-cap words, hashtags, and punctuation marks. The sentiment lexicon features are derived from lexicons described in Section 2. They include the number of sentiment words with non-zero sentiment score, the sum of sentiment scores of positive words (and separately negative words), and the sentiment score of the last token. We preprocess Arabic text by tokenizing with the CMU Twitter NLP tool to deal with specific tokens such as URLs, usernames, and emoticons. Then we use MADA to generate lemmas (Habash et al., 2009). Finally, we normalize different forms of Alif and Ya to bare Alif and dotless Ya.

We chose, for our experiments, an existing Arabic social media dataset—the BBN Arabic Dialectal Text (Zbib et al., 2012).[3] It contains sentences from the blog posts, with a mixture of expressions from the Levantine dialect of Arabic as well as Modern Standard Arabic. We randomly selected a subset of 1200 sentences, which we will refer to as the *BBN posts* or *BBN dataset*, and annotated them for sentiment on the crowdsourcing platform CrowdFlower.[4]

Table 3 shows ten-fold cross-validation accuracies obtained on the BBN dataset by our Arabic sentiment analysis system. Row a. shows results obtained using the various surface-form features as baseline. The rows within b. and c. show accuracies obtained by adding to the baseline system features derived from the Arabic sentiment lexicons and Arabic translations of English lexicons, respectively.

Observe that existing manually created sentiment lexicons (Abdul-Mageed et al. (2011) lexicon, Refaee and Rieser (2014) lexicon, and Kiritchenko et al. (2016) lexicon) provide only small improvements. While some of the automatically generated Arabic sentiment lexicons provide similar gains to the manual ones (accuracies around 63%), the Arabic Hashtag Lexicon (dialectal) helps obtain marked improvements in accuracy (65.3%).

Arabic translations of English sentiment lexicons also improve accuracies over the baseline, but not to the extent of the Dialectal Arabic Hashtag Lexicon. This is probably because the dialectal lexicon has dialectal Arabic terms, which are commonly used in social media texts. The dialectal lexicon also has a much larger set of terms than any of the manually created lexicons. Further, translation of English lexicon entries can lead to errors. The best results obtained with translations of English lexicons are from using the translated NRC Emotion Lexicon. Using both the Dialectal Hashtag Lexicon and the translated NRC Emotion Lexicon gives us the best results overall (66.6%).

[3] https://catalog.ldc.upenn.edu/LDC2012T09

[4] http://www.crowdflower.com

Resource	Number of instances			
	positive	negative	neutral	total
a. Arabic sentiment lexicons				
Manual lexicons:				
i. Abdul-Mageed et al. (2011) Lexicon	856	636	2490	3,982
ii. Refaee and Rieser (2014) Lexicon	135	348	-	483
iii. Kiritchenko et al. (2016) Lexicon	664	504	-	1,168

Table 1: Existing Arabic sentiment lexicons.

Resource	Number of instances			
	positive	negative	neutral	total
a. Arabic sentiment lexicons				
Automatic lexicons:				
i. Arabic Emoticon Lexicon	22,962	20,342	-	43,304
ii. Arabic Hashtag Lexicon	13,118	8,846	-	21,964
iii. Arabic Hashtag Lexicon (dialectal)	11,941	8,179	-	20,128
b. English lexicons translated into Arabic				
Manual lexicons:				
i. AFINN	878	1,598	-	2,476
ii. Bing Liu's Lexicon	2,006	4,783	-	6,789
iii. MPQA Subjectivity Lexicon	2,718	4,911	570	8,199
iv. NRC Emotion Lexicon	2,317	3,338	8,527	14,182
Automatic lexicons:				
v. NRC Emoticon Lexicon	15,210	11,530	-	26,740
vi. NRC Hashtag Lexicon	18,341	14,241	-	32,582

Table 2: New Arabic sentiment lexicons created as part of this project.

System	Accuracy (in percentage)
a. Baseline (uses word ngrams and other surface form features)	62.0
b. Baseline + Arabic lexicon	
Manual lexicons:	
i. Abdul-Mageed et al. (2011) Lexicon	62.2
ii. Refaee and Rieser (2014) Lexicon	63.0
iii. Kiritchenko et al. (2016) Lexicon	62.7
Automatic lexicons:	
iv. the Arabic Emoticon Lexicon	62.4
v. the Arabic Hashtag Lexicon	63.0
vi. the Arabic Hashtag Lexicon (dialectal)	**65.3**
vii. lexicon features from iv., v., and vi.	63.5
c. Baseline + Arabic translation of English lexicon	
Manual lexicons:	
i. English lexicon: AFINN	63.4
ii. English lexicon: Bing Liu Lexicon	63.0
iii. English lexicon: MPQA	61.9
iv. English lexicon: NRC Emotion Lexicon	**63.5**
Automatic lexicons:	
v. English lexicon: NRC Emoticon Lexicon	62.4
vi. English lexicon: NRC Hashtag Lexicon	61.7
d. Baseline + Arabic Hashtag Lexicon (dialectal) + Arabic translation of NRC Emotion Lexicon	**66.6**

Table 3: Sentiment classification accuracies on the BBN sentences. Highest scores in b., c., and d. are shown in bold.

	Before Translation	After Translation			
	# English Entries	# positive	# negative	# neutral	# changed
positive	100	85	9	6	15 (15.0%)
negative	100	4	92	4	8 (08.0%)
neutral	100	5	7	88	12 (12.0%)
All	300	94	108	98	35 (11.7%)

Table 4: Annotations of NRC Emotion Lexicon's sentiment association entries after automatic translation into Arabic.

Error categories	Percentage of total errors
1. Mistranslated	9.7
2. Translated to a related word	38.7
3. Translation correct, but 3a., 3b., or 3c.	51.6
3a. Different dominant sense	29.0
3b. Cultural differences	22.6
3c. Other reasons	0.0

Table 5: Percentage of erroneous entries assigned to each error category.

4. A Manual Study of Automatically Translated Sentiment Entries

As shown above, lexicons created by translating existing ones in other languages can be beneficial for automatic sentiment analysis. However, the above experiments do not explicitly quantify the extent to which such translated entries are appropriate, and how translation alters the sentiment of the source word. We conducted a small manual annotation study of 300 entries from the NRC Emotion Lexicon to determine the percentage of entries that were appropriate even after automatic translation into the focus language (Arabic). An appropriate entry is an Arabic translation that has the same sentiment association as its English source word. Additionally, translated entries that were deemed incorrect for Arabic were classified into coarse error categories. A list of pre-decided error categories was presented to the annotator, but the annotator was also encouraged to create new error categories as appropriate. The error categories provided are shown below:

1. The word is completely mistranslated.
2. The translation is not perfect, but the English word is translated into a word related to the correct translation. The Arabic word provided has a different sentiment than the English source word.
3. The translation is correct, but the Arabic word has a different sentiment than the English source word.
   (a) The dominant sense of the Arabic word is different from the dominant sense of the English source word, and they have different sentiments.
   (b) Cultural and life style differences between Arabic and English speakers lead to different sentiment associations of the English word and its translation.
   (c) Some other reason (give reason if you can).

The annotator was a native speaker of Arabic, who was also fluent in English.

We chose the NRC Emotion Lexicon for the study because it was manually created and because it has neutral terms as well (in addition to positive and negative terms). Since manual annotation is tedious, for this study, we randomly selected 100 positive words, 100 negative words, and 100 neutral words from the lexicon.

Table 4 shows the results of the human annotation study. Of the 100 positive entries examined, 85 entries were marked as appropriate in Arabic as well. Nine of the translations were marked as being negative in Arabic, and six were marked as neutral. Similarly, 92% of the translated negative entries and 88% of the translated neutral entries were marked appropriate in Arabic. Overall, 11.7% of the translated entries were deemed incorrect for Arabic.

Table 5 gives the percentage of erroneous entries assigned to each error category. Observe that close to 10% of the errors are caused by gross mistranslations, close to 40% of the errors are caused by translations into a related word, and about 50% of the errors are caused, not by bad translation, but by differences in how the word is used in Arabic—either because of different sense distributions (29%) or because of cultural differences (22.6%).

5. Conclusions

We created new Arabic sentiment lexicons using techniques of distant supervision, and showed their usefulness in the sentiment analysis of social media posts. We also translated existing English sentiment lexicons (four manually created ones and two that were created automatically) into Arabic using Google Translate. We showed that the lexicons improve performance over and above a competitive baseline classifier that uses various surface-form features. The Arabic Dialectal Hashtag Lexicon was especially useful, but adding features from translated lexicons further improved classification accuracy. Finally, we analyzed a subset of the automatically translated sentiment lexicon entries to show the extent to which sentiment is preserved after translation. We also identified the different reasons that can lead to erroneous entries in the translated lexicon. All of our lexicons are made freely available.

6. Bibliographical References

Abdul-Mageed, M., Diab, M. T., and Korayem, M. (2011). Subjectivity and sentiment analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 587–591, Portland, OR.

Badaro, G., Baly, R., Hajj, H., Habash, N., and El-Hajj, W. (2014). A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173, Doha, Qatar.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27.

Farra, N., Challita, E., Assi, R. A., and Hajj, H. (2010). Sentence-level and document-level sentiment mining for Arabic texts. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 1114–1119. IEEE.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA.

Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. (2014a). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August.

Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014b). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Kiritchenko, S., Mohammad, S. M., and Salameh, M. (2016). SemEval-2016 Task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, LA, California.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation Exercises*, SemEval '13, Atlanta, Georgia, USA, June.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016a). SemEval-2016 Task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016b). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '14, pages 27–35, Dublin, Ireland, August.

Refaee, E. and Rieser, V. (2014). An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015). SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June.

Salameh, M., Mohammad, S. M., and Kiritchenko, S. (2015). Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, Denver, Colorado.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA.

Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., and Ritter, A. (2013). SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59.

Zhu, X., Kiritchenko, S., and Mohammad, S. (2014). NRC-Canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland, August.