Arabic Sentiment analysis: an empirical study of machine translation's impact

Amira Barhoumi, Chafik Aloulou, Nathalie Camelin, Yannick Estève, Lamia Belguith

Arabic Sentiment analysis : an empirical study of machine translation's impact

Amira Barhoumi[1,2], Chafik Aloulou[2], Nathalie Camelin[1], Yannick Estève[1], and Lamia Hadrich Belguith[2]

[1] LIUM, Le Mans, France
amira.barhoumi.etu@univ-lemans.fr , nathalie.camelin@univ-lemans.fr , yannick.esteve@univ-lemans.fr
[2] MIRACL, Sfax University, Sfax, Tunisia
amirabarhoumi29@gmail.com , chafik.aloulou@fsegs.rnu.tn , l.belguith@fsegs.rnu.tn

Abstract. The largest amount of Sentiment Analysis has been carried out for English language. To deal with Arabic sentiment analysis, machine translation of English resources or Arabic texts may be applied to built Arabic sentiment analysis systems. In this paper, we translate Arabic dataset into English and study the impact of machine translation while considering a standard Arabic system as a baseline. Experiments show that sentiment analysis of Arabic content translated into English reach a competitive performance with respect to standard sentiment analysis of Arabic texts. This suggests that machine translation can successfully transfer the expression of sentiment or polarity. Moreover, we explored the multi-domain extending of training data in order to enhance performance and we show that we should have, in the training set, data whose domain is the same as the domain of evaluation dataset.

Keywords: Sentiment analysis, opinion mining, machine translation, document embeddings, machine learning, Arabic language.

Introduction

Sentiment analysis (SA) is a classification task that involves building systems recognizing the opinion expressed in natural language sentences. Its goal is commonly to identify the subjectivity (objective/subjective) and the polarity (positive/negative) of a given text [27].
Several research in SA have been carried out for English language. Nevertheless, there are few works that have been done for Arabic. This could be explained by the low number of resources developed in Arabic and their unavailability [3]. In sentiment analysis, making reliable resources is expensive and time-consuming because it needs experts to annotate corpora and built knowledge documents (lexicons, sentiment ontology, *etc.*). To avoid the cost of setting up resources, one track one may adopt is to translate existing resources from low-resources language to high-resources language and apply efficient methods for SA in that language.

Indeed, advances in machine translation (MT) have given Natural Language Processing (NLP) systems a strong boost. It was incorporated in many NLP applications such as online translation services, information extraction, document retrieval, *etc.* Research in opinion analysis took also advantage from machine translation, especially with under-resourced languages.
In this paper, we focus on the task of Arabic Sentiment Analysis (ASA) studying the impact of machine translation from Arabic to English on sentiment texts. Throughout a set of experiments presented in this paper, we answer several research questions such as:

1. Do MT systems alter the subjectivity or polarity expressed in the source Arabic text?
2. What kind of performances can be reached using MT systems in SA compared to standard SA on Arabic texts?
3. What kind of interest is there to use MT in ASA?
4. Does the use of a larger training data improve the results even if additional data comes from an other domain?

In this paper, we investigate how sentiment is preserved after machine translation. We conducted several experiments to study: the impact of MT systems; and the use of additional data from other domain in the training set. The remaining of this paper is organized as follow. In the next section, we give a short overview of ASA systems with a special attention to those using MT systems. Section 3 describes the methodology we propose to measure the impact of the use of MT in ASA. Section 4 presents the experimental setup, evaluation, performance and results discussion. Finally, we conclude and make suggestions for future researches in section 5 .

Related works

Research on sentiment analysis show a great interest from the scientific community as showed by various evaluation campaign [26], [22], [29], [23], [8]. If wide research has been carried out for English language, few works has been done for Arabic.

In this work, we focus on Arabic language. Indeed, many researchers have investigated sentiment analysis and opinion mining from different classification approaches. However, limited research is conducted on Arabic sentiment analysis. The research field on Arabic[3] is characterized by a luck on sentiment resources: annotated corpora and lexicons[4]. Thus, many research on ASA use machine translation to built Arabic resources thanks to the recent progress of MT systems especially from other languages into English.

In this section, we present some works that use machine translation to build Arabic Sentiment Analysis systems. In this framework, two tracks are possible:

[3] For an overview of Arabic sentiment analysis field, [1] build a survey.
[4] [3] and [9] summarizes all freely available corpora for Arabic sentiment analysis task

either translate from Arabic to English and train a system on English texts, or translate English resources into Arabic and train a system on Arabic texts. [30] and [24] tested the two tracks and they improve performance in sentiment prediction with both manual and machine translation documents. They conclude that sentiment analysis systems are able to capture polarity from translated texts. Moreover, [13] show that sentiment analysis of both original English dataset and the Arabic translated one produce comparable results. They infer that automatically translated English datasets shall be used as resources for building robust Arabic sentiment analysis systems.
It's commonly known that building SA system requires corpora and/or lexicon resources. Some researches ([28], [24]) translate Arabic corpora into English and others translate English lexicon into Arabic. In fact, [14] built an Arabic lexicon SLSA based on the translation of the English lexicon *SentiWordNet*[5] and Arabic morphological analyzer. In the same way, [6], [11] and [28] rely on the best state-of-the-art MT system and propose a method that leverages the available resource in another language such as the English *SentiWordNet*. Moreover, [15] exploited the machine translation to construct an Algerian sentiment lexicon (containing words in both Arabic and Arabizi). We can also mention [21] that used a seed list contained 14 English words , translated them to Arabic and filtered them based on the Arabic WordNet. [16] do the same by adding synonyms and translation to a basic lexicon. Moreover, [2] use both Arabic and English sentiment lexicons to classify the Arabic tweets into three sentiment categories (positive or negative or neutral).

The majority of works concludes that translation brings competitive results. However, [12] has shown that sentiment analysis accuracy does suffer when using translated lexicon and the quality of such lexicon is not as high as a manually constructed one.

[4] explores cross-lingual sentiment classification from English to Arabic, without any manual annotation effort, and found that it is easy to build and does not require deep linguistic analysis. They conclude that a good classification model can be obtained from translated corpora regardless of the noise added by machine translation.

Methodology

In this work, we are interested in studying the impact of machine translation on Arabic sentiment analysis. Thus, we propose the following method: to automatically translate Arabic text into English, to learn a classifier based on that translated text using English document embeddings as input. As a consequence, we can explore how English translation of Arabic text alters or not the expression and/or detection of sentiments.

In an other hand, we consider a standard ASA system trained on Arabic texts as a baseline. We will compare both systems performances.

The whole experimental scheme is outlined below:

[5] http://sentiwordnet.isti.cnr.it/

- Identify the arabic dataset *dataset_ar*.
- Learn an SA system on *dataset_ar* train based on pre-learned Arabic document embeddings. This system is denoted as the *ASA-baseline*.
- Run the ASA-baseline system on dataset_ar test and compute its performance.
- Translate *dataset_ar* from Arabic to English and obtain the English version *dataset_en*.
- Learn an SA system on dataset_en train based on pre-learned English document embeddings. This system is denoted *ASA-MT*.
- Run the ASA-MT classifier on on dataset_en test and compute its performance.
- Compare the performance of both systems ASA-baseline on dataset_ar and ASA-MT on dataset_en and draw some inferences.

Experiments and results

In this work, we choose to focus the sentiment analysis only considering the polarity expressed. As a consequence, the two SA systems implemented are binary classifiers that predict the two classes: *positive* and *negative*.

Architecture

We explore two classifiers: logistic regression (LR) and multi-layer perceptron (MLP)[6].

The input vector of each classifier is the embedding obtained by learning paragraph vector algorithm. It allows obtaining distributed representations (Doc2vec) for any length sequence, ranging from phrases to documents. It efficiently computes document vector representations in a dimensional vector space. The Doc2vec representations were used for English sentiment analysis by Le and Mikolov [19] who achieve the best performance with paragraph vector compared to other approaches on IMDB [20] dataset which contains 100000 film reviews.

Motivated by their work, we choose to use Doc2vec embeddings as input of our two systems. The input vector is a concatenation of two vectors: one learned from distributed memory version (DM) and one learned from distributed bag of words version (DBOW). Each one have 400 dimensions. So that, 800 is the dimension of the classifier input. As a result, we kept the same neural architecture and the same hyperparameters of paragraph vector model used in [19].

In order to translate Arabic texts into English, we used the LIUM[7] machine translation system. It is based on the statistical machine translation engine

[6] The MLP contains 3 layers: the input layer whose number of neurons is equal to the size of the input vector, one hidden layer with 50 units and the output layer with 2 neurons to predict the polarity of the input text: either positive or negative

[7] https://lium.univ-lemans.fr/

Moses[8][18]. It takes large quantities of parallel data (Arabic/English) and uses coocurrences of words and phrases to infer translation correspondences between the two languages. For decoding, *Moses* finds the highest scoring sentence in the target language (here, English).

Training data

The learning of Doc2vec representations needs a big corpus. According to our knowledge, LABR dataset [25] is the biggest arabic dataset for SA that is freely available[9].

Hence, we used LABR corpus for a set of various ASA experiments. This corpus consists of 63257 book reviews written in modern standard Arabic (MSA) and colloquial Arabic. Each review is associated by its author with a rate ranging from 1 to 5 stars. Table 1 describes the distribution of the reviews among the different ratings.

Table 1. Distribution of LABR dataset among rating stars.

	Very negative	Negative	Neutral	Positive	Very positive	Total
Training	2331	4195	9762	15189	19129	50606
Test	608	1090	2439	3865	1649	12651

It is customary to consider reviews with 1 or 2 stars as *negative* reviews and those with 4 or 5 stars as *positive*. Reviews with 3 stars are neutral and they are not considered in classification. Thus, final corpus is reduced to 40845 reviews (68% positive) for the training corpus and 10211 for the test corpus (69% positive). Note that 10% of the training set is used as a development corpus.

Feature extraction and experimental setup

As mentioned in the architecture sub-section, two classifiers were investigated: logistic regression and multi-layer perceptron. The input of each classifier is a set of document embeddings.

We tested three different types of document embedding: DM, DBOW and the concatenation DM+DBOW. We trained the classifiers with different learning rates (10^{-3}, 10^{-4} et 10^{-5}). While varying different hyper-parameters, LR classifier gives better results than MLP. That is why we only reported in this work the performances of logistic regression, MLP results do not inferred more interesting information.

[8] http://www.statmt.org/moses/
[9] LARB dataset is available on http://www.mohamedaly.info/datasets/labr

To evaluate the performance of SA on the LABR dataset, we carried out several experiments using various configurations. All the experiments were conducted in Python using Theano[10] for classification and Gensim[11] for learning vector representation. The performance is measured with *error rate*[12] metric which calculates the percentage of misclassified examples.

Results and discussion

First, the two system ASA-baseline and ASA-MT were running as detailed in previous sub-sections. Table 2 shows the error rates of both systems obtained with the logistic regression classifier.

Table 2. Error rates of Logistic regression classifier over Arabic and English-translated datasets: LABR_ar and LABR_en respectively.

System	Dataset input	Error rate
ASA-baseline	LABR_ar	25.37%
ASA-MT	LABR_en	23.70%

The error rate of ASA-baseline system is 25.37% that drops to 23.7% when using the ASA-MT system. So, we note a gain of 1.7% with the proposed ASA-MT system. This latter is trained on LABR_en dataset which represent the English translation of LABR_ar used in the ASA-baseline. So that, we highlight that machine translation does not seem to alter the expression of a polarity and brings a competitive (even better) accuracy with respect to Arabic sentiment analysis.

In order to explain and understand why translating the arabic LABR dataset in English enhance the performance, we analyzed the translated corpus LABR_en. One transformation done during the translation process is to remove all non-translated words from LABR_en as these words may be ambiguous or noisy in the English dataset. Considering that choice and the observed results, we have made the assumption that non-translated words may carry ambiguity. To validate this hypothesis, we conducted a non-translated Arabic word's analysis. It shows that these non-translated words are mainly: dialectal words, proper names, typo words, words with repetition of characters, non-Arabic origin words written with Arabic letters (such as: الرفيو /Alrfyw/ for *review*, البروتكشن /Albrwtk$n/for *protection*, *etc.*). Following this analysis, we consider that these non-translated

[10] http://deeplearning.net/software/theano/
[11] https://radimrehurek.com/gensim/
[12] Error rate = 1 - Accuracy

words seem to disrupt the polarity detection, so we choose to consider these words as noisy ones. In order to see if sufficient information is kept in the English corpus to extract the polarity, we choose to remove the noisy words from the Arabic dataset. This corpus is denoted LABR_ar_MTremove. A new LR classifier is learned on LABR_ar_MTremove. We hope that this new classifier would have a positive impact on performance. Unfortunately, we obtained an error rate of 26.86% in that configuration which is greater than 25.37% initially obtained on LABR_ar. In that way, this removal words seems to be informative for the standard ASA-baseline system, even if they are not for the MT-system.

Other ASA systems using NLP preprocess tools were implemented like [7] that chose to apply light stemming as a pre-processing step on LABR_ar. As a result, they obtained an error rate of 23.31% which is better than 25.37% obtained by our ASA-baseline. This means that light stemming is a reliable pre-process step for Arabic sentiment analysis. Moreover, it is higher (approximatly equal) than 23.7% obtained on the English dataset LABR_en. We infer that machine translation, as a statistic tool, and light stemming, as a linguistic tool, almost behave the same way for the Arabic sentiment analysis task. So for languages without such linguistic tools, we could apply machine translation to have a good sentiment analysis system, especially with the progress of machine translation field. In other words, we could use machine translation if linguistic specific tool is not yet developed or not powerful.

To go further to explore the impact of MT systems in ASA, we believe that the performance obtained with the proposed ASA-MT system may be enhanced using a larger amount of training data. Moreover, we believe that a larger corpus of reviews, whatever the domain, should be positive. In this perspective, we made some experiments while varying the domain (books, films) of reviews. This setup is as follows:

- Training set composed of IMDB dataset [20].
- Training set composed of IMDB dataset and the LABR training set.
- Training set composed of LABR training set and the classifier's parameters are initialized with the parameters obtained after training the classifier on IMDB dataset.

Results are reported in Table 3. We can observe that, by changing the domain of the training set, the error rate increases from 23.7% to 29.33%. However, while adding reviews of other domain to the training set, we obtain 26.94% as error rate. So, it performs better than changing completely the training domain (26.94% vs. 29.33%).
Other than mixing data with different domains, multi-domain experiment can be also established by initializing the classifier parameters with those obtained while training on dataset of other domain. The error rate reaches 24.84% with this technique. It is better than training on domain-mixed data (24.84% vs. 26.94%).

Table 3. Experiment's configurations to enlarge the training set.

	Partition sets	Error rate
Domain change	Train = IMDB Dev = LABR-dev Test = LABR-test	29.33%
Multi-Domain	Train = IMDB + LABR-train Dev = LABR-dev Test = LABR-test	26.94%
	Train = LABR-train Dev = LABR-dev Test = LABR-test	24.84%

Conclusion and future works

In this paper, we presented a set of experiments to study the impact of English machine translation on sentiment analysis of Arabic reviews. Our experiments show that sentiment analysis of English translation is better than sentiment analysis of gross Arabic texts. We observed that machine translation does not alter the polarity prediction. Moreover, we found that sentiment analysis of English translation reach competitive results with respect to sentiment analysis of light-stemmed Arabic texts. So, we could generalize, whatever the language, that machine translation could be used if such linguistic tools (light stemming) do not exist or are not efficient. We have also explored the track of extending the training corpus and we proved the interest of keeping data whose domain is the same as the test set domain. We also showed that mixing the training data domains performs better than simply changing the training domain.

As future work, we think combining light stemming and machine translation based systems. In fact, these systems behave differently as they use different techniques. We strongly think that combining them could enhance performances.

Moreover, one other perspective is to investigate the use of deep learning classifiers, especially convolutional neural networks (CNN) which are efficient in English sentiment analysis ([17], [5], [32],[10]). We think testing CNN and other deep learning techniques on ASA.

Finally, we would explore different types of embedding, in addition to document embeddings, like sentiment-specific word embeddings [31].

References

1. Al-Ayyoub, M., Khamaiseh, A.A., Jararweh, Y., Al-Kabi, M.N.: A comprehensive survey of arabic sentiment analysis. Information Processing & Management (2018)
2. Al-Horaibi, L., Khan, M.B.: Sentiment analysis of arabic tweets using semantic resources. International Journal of Computing & Information Sciences **12**(2), 149 (2016)
3. Al-Kabi, M., Al-Ayyoub, M., Alsmadi, I., Wahsheh, H.: A prototype for a standard arabic sentiment analysis corpus. Int. Arab J. Inf. Technol. **13**(1A), 163–170 (2016)

4. Al-Shabi, A., Adel, A., Omar, N., Al-Moslmi, T.: Cross-lingual sentiment classification from english to arabic using machine translation. International journal of advanced computer science and applications **8**(12), 434–440 (2017)
5. Alayba, A.M., Palade, V., England, M., Iqbal, R.: A combined cnn and lstm model for arabic sentiment analysis. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 179–191. Springer (2018)
6. Alotaibi, S.S., Anderson, C.W.: Extending the knowledge of the arabic sentiment classification using a foreign external lexical source. Int. J. Nat. Lang. Comput **5**(3), 1–11 (2016)
7. Barhoumi, A., Estève, Y., Aloulou, C., Hadrich Belguith, L.: Document embeddings for arabic sentiment analysis. Language Processing and Knowledge Management (2017)
8. Benamara, F., Grouin, C., Karoui, J., Moriceau, V., Robba, I.: Analyse d'opinion et langage figuratif dans des tweets présentation et résultats du défi fouille de textes DEFT2017. In: Actes de DEFT. Orléans, France (2017)
9. Boudad, N., Faizi, R., Oulad Haj Thami, R., Chiheb, R.: Sentiment analysis in arabic: A review of the literature. Ain Shams Eng J (2017), http://dx.doi.org/10.1016/j.asej.2017.04.007 (2017)
10. Croce, D., Castellucci, G., Basili, R.: Injecting sentiment information in context-aware convolutional neural networks. In: IIR (2016)
11. Duwairi, R., Ahmed, N.A., Al-Rifai, S.Y.: Detecting sentiment embedded in arabic social media–a lexicon-based approach. Journal of Intelligent & Fuzzy Systems **29**(1), 107–117 (2015)
12. El-Beltagy, S.R.: Weightednileulex: A scored arabic sentiment lexicon for improved sentiment analysis. Language Processing, Pattern Recognition and Intelligent Systems. Special Issue on Computational Linguistics, Speech& Image Processing for Arabic Language. World Scientific Publishing Co (2017)
13. Elnagar, A., Einea, O., Lulu, L.: Comparative study of sentiment classification for automated translated latin reviews into arabic. In: 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA). pp. 443–448 (2017). https://doi.org/10.1109/AICCSA.2017.82
14. Eskander, R., Rambow, O.: Slsa: A sentiment lexicon for standard arabic. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2545–2550 (2015)
15. Guellil, I., Adeel, A., Azouaou, F., Hussain, A.: Sentialg: Automated corpus annotation for algerian sentiment analysis. arXiv preprint arXiv:1808.05079 (2018)
16. Ibrahim, H.S., Abdou, S.M., Gheith, M.: Automatic expandable large-scale sentiment lexicon of modern standard arabic and colloquial. In: Arabic Computational Linguistics (ACLing), 2015 First International Conference on. pp. 94–99. IEEE (2015)
17. Kim, Y.: Convolutional neural networks for sentence classification. In: In EMNLP. Citeseer (2014)
18. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. pp. 177–180. Association for Computational Linguistics (2007)
19. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1188–II–1196. ICML'14, JMLR.org (2014), http://dl.acm.org/citation.cfm?id=3044805.3045025

20. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 142–150. Association for Computational Linguistics (2011)
21. Mahyoub, F.H., Siddiqui, M.A., Dahab, M.Y.: Building an arabic sentiment lexicon using semi-supervised learning. Journal of King Saud University-Computer and Information Sciences **26**(4), 417–424 (2014)
22. May, J.: Semeval-2016 task 8: Meaning representation parsing. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 1063–1073. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/S16-1166
23. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 task 1: Affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation. pp. 1–17. Association for Computational Linguistics (2018), http://aclweb.org/anthology/S18-1001
24. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How translation alters sentiment. Journal of Artificial Intelligence Research **55**, 95–130 (2016)
25. Nabil, M., Aly, M., Atiya, A.: Labr: A large scale arabic sentiment analysis benchmark. arXiv preprint arXiv:1411.6718 (2014)
26. Nakov, P., Zesch, T., Cer, D., Jurgens, D. (eds.): Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, Denver, Colorado (June 2015), http://www.aclweb.org/anthology/S15-2
27. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval **2**(1–2), 1–135 (2008)
28. Refaee, E., Rieser, V.: Benchmarking machine translated sentiment analysis for arabic tweets. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. pp. 71–78 (2015)
29. Rosenthal, S., Farra, N., Nakov, P.: SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation. SemEval '17, Association for Computational Linguistics, Vancouver, Canada (August 2017)
30. Salameh, M., Mohammad, S., Kiritchenko, S.: 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. pp. 767–777 (2015)
31. Yu, L.C., Wang, J., Lai, K.R., Zhang, X.: Refining word embeddings for sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 534–539 (2017)
32. Zhang, R., Lee, H., Radev, D.: Dependency sensitive convolutional neural networks for modeling sentences and documents. In: Proceedings of NAACL-HLT. pp. 1512–1521 (2016)