Sentiment after Translation: A Case-Study on Arabic Social Media Posts

Mohammad Salameh
University of Alberta
msalameh@ualberta.ca

Saif M. Mohammad and Svetlana Kiritchenko
National Research Council Canada
{saif.mohammad,svetlana.kiritchenko}
@nrc-cnrc.gc.ca

Abstract

When text is translated from one language into another, sentiment is preserved to varying degrees. In this paper, we use Arabic social media posts as stand-in for source language text, and determine loss in sentiment predictability when they are translated into English, manually and automatically. As benchmarks, we use manually and automatically determined sentiment labels of the Arabic texts. We show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. We discover that even though translation significantly reduces the human ability to recover sentiment, automatic sentiment systems are still able to capture sentiment information from the translations.

1 Introduction

Automatic sentiment analysis of text, especially social media posts, has a number of applications in commerce, public health, and public policy development. However, a vast majority of prior research on automatic sentiment analysis has been on English texts. Furthermore, many sentiment resources essential to automatic sentiment analysis (e.g., sentiment lexicons) exist only in English. Thus there is a growing need for effective methods for analyzing text from other languages such as Arabic and Chinese, especially posts on social media. There has also been marked progress in automatic translation of texts, especially from other languages into English. Thus, instead of building source-language specific sentiment analysis systems, one can translate the texts into English and use an English sentiment analysis system. However, it is widely believed that aspects of sentiment may be lost in translation, especially in automatic translation. Though, the extent of this loss, in terms of drop in accuracy of automatic sentiment systems remains undetermined.

This paper analyzes several methods available in annotating non-English texts for sentiment:

- Use a source-language sentiment analysis system.
- Run an English sentiment analysis system on manually created English translations of source language text.
- Run an English sentiment analysis system on automatically generated English translations of source language text.

In our experiments, we use Arabic social media posts as a specific instance of the source language text. We use state-of-the-art Arabic and English sentiment analysis systems as well as a state-of-the-art Arabic-to-English translation system. We outline the advantages and disadvantages of each of the methods listed above, and more importantly conduct experiments to determine accuracy of sentiment labels obtained using each of these methods. As benchmarks we use manually and automatically determined sentiment labels of the Arabic tweets.

These results will help users determine methods best suited for their particular needs. Along the way, we answer several research questions such as:

1. What sentiment prediction accuracy is expected when Arabic blog posts and tweets are

translated into English (using the current state-of-art techniques), and then run through a state-of-the-art English sentiment analysis system?

2. How does this performance compare with that of a current state-of-the-art Arabic sentiment system?
3. What is the loss in sentiment predictability when translating Arabic text into English automatically vs. manually?
4. How difficult is it for humans to determine sentiment of automatically translated text?
5. When dealing with translated text, which is more accurate at determining the sentiment of Arabic text: (1) automatic sentiment analysis of the translated text, or (2) human annotation of the translated text for sentiment?

The inferences drawn from these experiments do not necessarily apply to language pairs other than Arabic–English. Languages can differ significantly in terms of characteristics that impact accuracy of an automatic sentiment analysis system. Our goal here specifically is to understand sentiment predictability of Arabic dialectal text on translation. However, a similar set of experiments can be used for other language pairs as well to determine the impact of translation on sentiment.

Through our experiments on two different datasets, we show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. We also show that translation (both manual and automatic) introduces marked changes in sentiment carried by the text; positive and negative texts can often be translated into texts that are neutral. We also find that certain attributes of automatically translated text that mislead humans with regards to the true sentiment of the source text, do not seem to affect the automatic sentiment analysis system.

In the process of developing these experiments to study how translation alters sentiment, we created a state-of-the-art Arabic sentiment analysis system by porting NRC-Canada's competition winning system (Kiritchenko et al., 2014) to Arabic. We also created a substantial amount of sentiment labeled data pertaining to Arabic social media texts and their English translations which is made freely available.[1] This is the first such resource where text in one language and its translations into another language (both manually and automatically produced) are each manually labeled for sentiment.

[1] http://www.purl.com/net/ArabicSentiment

2 Related Work

2.1 Sentiment Analysis of English Social Media

Sentiment analysis systems have been applied to many different kinds of texts including customer reviews, newspaper headlines (Bellegarda, 2010), novels (Boucouvalas, 2002; Mohammad and Yang, 2011), emails (Liu et al., 2003; Mohammad and Yang, 2011), blogs (Neviarouskaya et al., 2011), and tweets (Mohammad, 2012). Often these systems have to cater to the specific needs of the text such as formality versus informality, length of utterances, etc. Sentiment analysis systems developed specifically for tweets include those by Go et al. (2009), Pak and Paroubek (2010), Agarwal et al. (2011), and Thelwall et al. (2011). A survey by Martínez-Cámara et al. (2012) provides an overview of the research on sentiment analysis of tweets. In the last two years, several shared tasks on sentiment analysis were organized by the Conference on Semantic Evaluation Exercises (SemEval), which allowed for comparison of different approaches on common datasets from different domains (Wilson et al., 2013; Rosenthal et al., 2014; Pontiki et al., 2014). The NRC-Canada system (Kiritchenko et al., 2014) ranked first in these competitions, and we use it in our experiments. Details of the system are described in Section 6.

2.2 Sentiment Analysis of Arabic Social Media

Sentiment analysis of Arabic social media texts has several challenges. The text is often in a regional Arabic dialect rather than Modern Standard Arabic (MSA). Unlike MSA which is a standardized form of Arabic, dialectal Arabic is the spoken form of Arabic and lacks strict writing standards. The text often includes words from languages other than Arabic and multiple scripts may be used to express Arabic and foreign words. In addition, Arabic is a morphologically complex language, thus having a lexicon of word-sentiment associations that covers all different surface forms becomes a cumbersome task. Negation in MSA is expressed through negation par-

ticles, but in some dialects (Egyptian) it is expressed using suffixes at the end of the word. We refer the reader to Mourad and Darwish (2013) for more details on these issues.

There have been a few studies tackling sentiment analysis of Arabic texts (Ahmad et al., 2006; Badaro et al., 2014). The ones most closely related to our work are the studies of sentiment analysis of Arabic social media (Al-Kabi et al., 2013; El-Beltagy and Ali, 2013; Mourad and Darwish, 2013; Abdul-Mageed et al., 2014). Here we review existing Arabic sentiment analysis systems that were designed specifically for Arabic social media datasets. Abdul-Mageed et al. (2014) trained an SVM classifier on a manually labeled dataset and applied a two-stage classification that first separates subjective from objective sentences and then classifies the subjective into positive or negative instances. The authors have compiled several datasets from multiple social media resources that include chatroom messages, tweets, forum posts, and Wikipedia Talk pages. However, these resources have not been made publicly available yet.

Mourad and Darwish (2013) trained SVM and Naive Bayes classifiers on Arabic tweets annotated by two native Arabic speakers. We compare our system's performance to theirs in Section 7.

Refaee and Rieser (2014b) manually annotated tweets for sentiment by two native Arabic speakers. They used an SVM to classify tweets in a two-stage approach, polar vs neutral, then positive vs. negative. The authors shared their data with us and we test our system on their dataset. However, the dataset they provided us is a larger superset than the one they had originally used (Refaee and Rieser, 2014a). Thus, the results of sentiment systems on the two sets are not directly comparable.

2.3 Multilingual Sentiment Analysis

Work on multilingual sentiment analysis has mainly addressed mapping sentiment resources from English into morphologically complex languages. Mihalcea et al. (2007) used English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon is then used to classify Romanian text. Wan (2008) translated Chinese customer reviews to English using a machine translation system. The translated reviews are then classified with a rule-based system that relies on English lexicons. A higher accuracy is achieved by using ensemble methods and combining knowledge from Chinese and English resources. Balahur and Turchi (2014) conducted a study to assess the performance of statistical sentiment analysis techniques on machine-translated texts. Opinion-bearing phrases from the *New York Times* text corpus (2002–2005) were automatically translated using publicly available machine-translation engines (Google, Bing, and Moses). Then, the accuracy of a sentiment analysis system trained on original English texts was compared to the accuracy of the system trained on automatic translations to German, Spanish, and French. The authors concluded that the quality of machine translation is sufficient for sentiment analysis to be performed on automatically translated texts without a substantial loss in accuracy. Contrary to that work, our study uses both manual and automatic translations as well as both manual and automatic sentiment assignments to systematically examine the effect of translation on sentiment. Additionally, we deal with noisy social media texts as opposed to more polished news media texts. There exists research on using sentiment analysis to improve machine translation (Chen and Zhu, 2014), but that is beyond the scope of this paper.

3 Method for Determining Sentiment Predictability on Translation

In order to systematically study the impact of translation on sentiment analysis, we propose the following experimental setup:

- Identify or compile an Arabic social media dataset. We will refer to it as *Ar*. (*Ar* comes from the first two letter of Arabic.)
- Manually translate *Ar* into English. We will refer to these English translations as *En(Manl.Trans.)* [*Manl.* is for manual, and *Trans.* is for translations.]
- Automatically translate *Ar* into English. We will refer to these English translations as *En(Auto.Trans.)* [*Auto.* is for automatic.]
- Manually annotate Ar. for sentiment. We will refer to the sentiment-labeled dataset as *Ar(Manl.Sent.)*

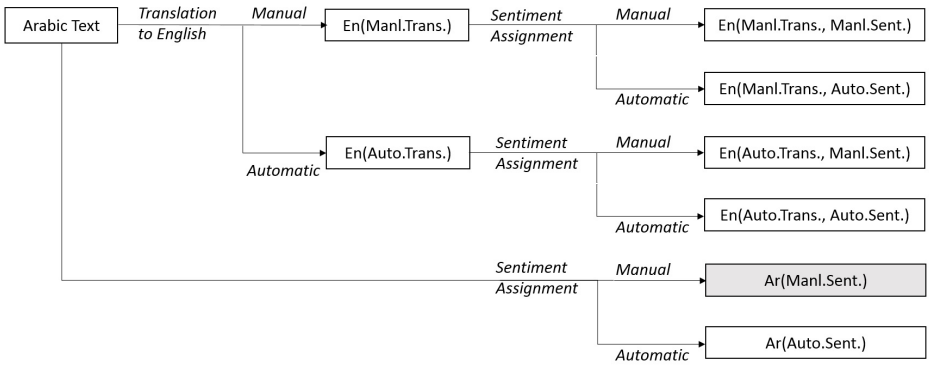

Figure 1: Experimental setup to determine the impact of translation on sentiment. We compare sentiment labels between Ar(Manl.Sent.) (shown in a shaded box) and other datasets shown on the right side of the figure. Ar(Manl.Sent.) is the original Arabic text manually annotated for sentiment.

- Manually annotate all English datasets [*En(Manl.Trans.)* and *En(Auto.Trans.)*] for sentiment, creating *En(Manl.Trans., Manl.Sent.)* and *En(Auto.Trans., Manl.Sent.)*, respectively.
- Run a state-of-the-art Arabic sentiment analysis system on *Ar*, creating *Ar(Auto.Sent.)*
- Run a state-of-the-art English sentiment analysis system on all the English datasets [*En(Manl.Trans.)* and *En(Auto.Trans.)*], creating *En(Manl.Trans., Auto.Sent.)* and *En(Auto.Trans., Auto.Sent.)*, respectively.

Figure 1 depicts this setup. Once the various sentiment-labeled datasets are created, we can compare pairs of datasets to draw inferences. For example, comparing the labels for *Ar(Manl.Sent.)* and *En(Manl.Trans., Manl.Sent.)* will show how different the sentiment labels tend to be when text is translated from Arabic to English. The comparison will also show, for example, whether positive tweets tend to often be translated into neutral tweets, and to what extent. The results will also show how feasible it is to first translate Arabic text into English and then use automatic sentiment analysis (*Ar(Manl.Sent.)* vs. *En(Auto.Trans., Auto.Sent.)*). In Section 8, we provide an analysis of several such comparisons for two different Arabic social media datasets.

DATA: Since manual translation of text from Arabic to English is a costly exercise, we chose, for our experiments, an existing Arabic social media dataset that has already been translated – the BBN Arabic-Dialect/English Parallel Text (Zbib et al., 2012).[2] It contains about 3.5 million tokens of Arabic dialect sentences and their English translations. We use a randomly chosen subset of 1200 Levantine dialectal sentences, which we will refer to as the *BBN posts* or *BBN dataset*, in our experiments. Additionally, we also conduct experiments on a dataset of 2000 tweets originating from Syria (a country where Levantine dialectal Arabic is commonly spoken). These tweets were collected in May 2014 by polling the Twitter API. We will refer to this dataset as the *Syrian tweets* or *Syrian dataset*. Note, however, that manual translations of the Syrian dataset are not available.

The experimental setup described above involves several component tasks: generating translations manually and automatically (Section 4), manually annotating Arabic and English texts for sentiment (Section 5), automatic sentiment analysis of English texts (Section 6), and automatic sentiment analysis of Arabic texts (Section 7).

4 Generating English Translations

The BBN dialectal Arabic dataset comes with manual translations into English. We generate automatic translations of the Arabic BBN posts and the Syrian tweets, by training a multi-stack phrase-based machine translation system to translate from Arabic to English. Our in-house system is quite similar to Cherry and Foster (2012). This statistical machine translation (SMT) system is trained on data from OpenMT 2012. We preprocess the training data by

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

segmenting the Arabic source side of the training data with MADA 3.2 (Habash et al., 2009), using Penn Arabic Treebank (PATB) segmentation scheme as recommended by El Kholy and Habash (2012). The Arabic script is further normalized by converting different forms of Alif إ أ آ ا and Ya ى ي to bare Alif ا and dotless Ya ى. The different forms are used interchangeably, and normalization decreases the sparcity of Arabic tokens and improves translation. The English side of the training data is lowercased and tokenized by stripping punctuation marks. We set the decoder's stack size to 10000 and distortion limit to 7. We replace the *out-of-vocabulary* words in the translated text with *UNKNOWN* token (which is shown to the annotators). The decoder's log-linear model is tuned with MIRA (Chiang et al., 2008; Cherry and Foster, 2012). A KN-smoothed 5-gram language model is trained on the English Gigaword and the target side of the parallel data.

5 Creating sentiment labeled data in Arabic and English

Manual sentiment annotations were performed on the crowdsourcing platform CrowdFlower[3] for three BBN datasets and two Syrian datasets:

1. Original Arabic posts (BBN and Syria datasets), annotated by Arabic speakers.
2. Manual English translations of Arabic posts, annotated by English speakers (only for BBN dataset).
3. Automatic English translations of Arabic posts (BBN and Syria datasets), annotated by English speakers.

Each post was annotated by at least ten annotators and the majority sentiment label was chosen. Table 1 shows the class distribution of sentiment labels in various datasets. Observe from rows *a* and *d* that neutral tweets constitute only about 10% of the data in both BBN and Syria datasets. The Syrian tweets have a much higher percentage of negative posts, whereas in the BBN data, the percentages of positive and negative posts are comparable. (Arabic tweets in general tend to be much more skewed to the negative class than Arabic blog post sentences.) Rows b, c, and e show that translated texts tend to lose some of the sentiment information and there is a relatively higher percentage of neutral instances in the translated text than in the original text.

For each post, we determine the count of the most frequent annotation divided by the total number of annotations. This score is averaged for all posts to determine the inter-annotator agreement shown in the last column of Table 1. We use this agreement score as benchmark to compare performance of automatic sentiment systems (described below).

6 English Sentiment Analysis

We use the English-language sentiment analysis system developed by NRC-Canada (Kiritchenko et al., 2014) in our experiments. This system obtained highest scores in two recent international competitions on sentiment analysis of tweets –SemEval-2013 Task 2 and SemEval-2014 Task 9 (Wilson et al., 2013; Rosenthal et al., 2014). We briefly describe the system below; for more details, we refer the reader to Kiritchenko et al. (2014).

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, semantic, and sentiment lexicon features described below. The sentiment lexicon features are derived from existing, general-purpose, manual lexicons, namely NRC Emotion Lexicon (Mohammad and Turney, 2010; Mohammad and Turney, 2013), Bing Liu's Lexicon (Hu and Liu, 2004), and MPQA Subjectivity Lexicon (Wilson et al., 2005), as well as automatically generated, tweet-specific lexicons, Hashtag Sentiment Lexicon and Sentiment140 Lexicon (Kiritchenko et al., 2014).[4]

6.1 Generating English Sentiment Lexicon

Ablation experiments in Mohammad et al. (2013) showed that their sentiment system benefited most from the use of the Hashtag Sentiment Lexicon. The lexicon was created as follows. A list of 77 seed words, which are synonyms of *positive* and *negative*, was compiled from the Roget's Thesaurus. Then, the Twitter API was polled to collect tweets that had these words as hashtags. A tweet is considered positive if it has a positive hashtag and negative if it

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

[4] http://www.purl.com/net/lexicons

	positive	negative	neutral	agreement
BBN data				
a. Ar(Manl.Sent)	41.50	47.92	10.58	73.80
b. En(Manl.Trans., Manl.Sent)	35.00	43.25	21.75	68.00
c. En(Auto.Trans., Manl.Sent)	36.17	36.50	27.34	65.70
Syria data				
d. Ar(Manl.Sent)	22.40	67.50	10.10	79.00
e. En(Auto.Trans., Manl.Sent)	14.25	66.15	19.60	76.10

Table 1: Class distribution (in percentage) of the sentiment annotated datasets.

has a negative hashtag. For each term in the tweet set, a sentiment score is computed by measuring the PMI (pointwise mutual information) between the term and the positive and negative categories:

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

where w is a term in the lexicon. $PMI(w, pos)$ is the PMI score between w and the positive class, and $PMI(w, neg)$ is the PMI score between w and the negative class. A positive *SenScore* (w) suggests that the word is associated with positive sentiment and a negative score suggests that the word is associated with negative sentiment. The magnitude indicates the strength of the association.

6.2 Pre-processing and Feature Generation

The following pre-processing steps are performed. URLs and user mentions are normalized to http://someurl and @someuser, respectively. Tweets are tokenized and part-of-speech tagged with the CMU Twitter NLP tool (Gimpel et al., 2011). Then, each tweet is represented as a feature vector.

The features:
- Word and character ngrams;
- POS: # occurrences of each part-of-speech tag;
- Negation: # negated contexts. Negation also affects the ngram features: a word w becomes w_NEG in a negated context;
- Automatic sentiment lexicons: For each token w occurring in a tweet, its sentiment score $score(w)$ is used to compute: # tokens with $score(w) \neq 0$; the total score = $\sum_{w \in tweet} score(w)$; the maximal score = $max_{w \in tweet} score(w)$; the score of the last token in the tweet.
- Manually created sentiment lexicons: For each of the three manual sentiment lexicons, the following features are computed: the sum of positive and the sum of negative scores for tweet tokens in affirmative contexts and in negated contexts, separately.

7 Arabic Sentiment Analysis

7.1 Building an Arabic Sentiment System

We built an Arabic sentiment analysis system by reconstructing the NRC-Canada English system to deal with Arabic text. It extracts all of the feature described in Section 6.2 except POS and negation features. We also generated a word-sentiment association lexicon as described in Section 6.1, but for Arabic words from Arabic tweets (more details in sub-section below). We preprocess Arabic text by tokenizing with CMU Twitter NLP tool to deal with specific tokens such as URLs, usernames, and emoticons. Then we use MADA to generate lemmas. Finally, we normalize different forms of Alif and Ya to bare Alif and dotless Ya to decrease token sparcity in Arabic datasets.

7.1.1 Generating Arabic Sentiment Lexicon

We translated 77 positive and negative seed words used to generate the English NRC Hashtag Sentiment Lexicon into Arabic using Google Translate. Among the several translations provided by it, we chose words that were less ambiguous and tended to have strong sentiment in Arabic texts. To increase the coverage of our seed list, we manually added different inflections for these translations.

We polled the Twitter API for the period of June to August 2014 and collected tweets with *#(keyword)*. After filtering out duplicate tweets and retweets, we ended up with 163,944 positive unique tweets and 37,848 negative unique tweets. Then for each word w, *SenScore* (w) was calculated just as described in Section 6.1.

Arabic Sentiment Labeled Dataset	**MD**	**RR**	**BBN**	**Syria**
sentiment classes	pos, neg	pos,neg	pos, neg, neu	pos, neg, neu
number of instances	1111	2681	1199	2000
Most frequent class baseline	66.06	68.92	47.95	67.50
Human agreement benchmark	-	-	73.82	79.05
Mourad and Darwish Arabic SA system	72.50	-	-	-
Our Arabic SA system	74.62	85.23	63.89	78.65

Table 2: Accuracy (in percentage) of sentiment analysis (SA) systems on various Arabic social media datasets.

	pos	**neg**	**neu**
BBN data			
a. Ar(Auto.Sent)	39.78	60.05	0.17
b. En(Manl.Trans., Auto.Sent)	43.12	55.63	1.25
c. En(Auto.Trans., Auto.Sent)	42.87	56.05	1.08
Syria data			
d. Ar(Auto.Sent)	20.60	75.30	4.10
e. En(Auto.Trans., Auto.Sent)	24.75	69.75	5.50

Table 3: Class distribution (in percentage) resulting from automatic sentiment analysis.

7.2 Evaluation

We tested the Arabic sentiment system on two existing Arabic datasets (Mourad and Darwish (2013) (MD) and Refaee and Rieser (2014a) (RR)) and two newly sentiment-annotated Arabic datasets (BBN and Syria). Table 2 shows results of ten-fold cross-validation experiments on each of the datasets. For MD and RR, the presented results are for the two-class problem (positive vs. negative) to allow for comparison with prior published results. For BBN and Syria, the results are shown for the case where the system has to identify one of three classes: positive, negative, or neutral. Human agreement scores are shown where available.

Note that the accuracy of our system is higher than previously published results on the MD dataset. The only previously published results on the RR dataset are on a small subset (about 1000 instances) for which Refaee and Rieser (2014a) obtained an accuracy of 87%. The results in Table 2 are for a larger dataset and so not directly comparable.

8 Sentiment After Translation

Using the methods and systems described in Sections 4, 5, 6, and 7, we generated all the manually and automatically labeled datasets mentioned in Section 3's Experimental Setup. Table 3 shows the distribution of positive, negative, and neutral classes in datasets that have been automatically labeled with sentiment. These percentages can be compared with those in Table 1 (rows a and d) which show the true sentiment distribution in the BBN and Syria datasets. Observe that the automatic system has difficulty in assigning neutral class to posts. This is probably because of the small percentage (about 10%) of neutral tweets in the training data. Also notice that the system predominantly guesses negative, which is also a reflection of the distribution in the training data. The strong bias to negatives is lessened in the English translations.

Main Result: Tables 4 and 5 show how similar the sentiment labels are across various pairs of datasets for the BBN posts and the Syrian posts, respectively. For example, row a. in Table 4 shows the comparison between Arabic tweets that were manually annotated for sentiment and those that were automatically labeled for sentiment by our Arabic sentiment analysis system. Column 2 shows the percentage of instances where the sentiment labels match across the two datasets being compared. For row a. the match percentage of 63.89% represents the accuracy of the automatic sentiment analysis system on the Arabic BBN posts.

Row b. shows the difference in labels when text is manually translated from Arabic to English, even though sentiment labeling in both Arabic and English is done manually. Observe that the two labels match only 71.31% of the time. However, the agreement among human sentiment annotators on original Arabic texts was only 73.8%. So, the English translation does affect sentiment, but not dramatically.

Row c. shows results for when the manually translated text is run through an English sentiment analysis system and the labels are compared against *Ar(Manl.Sent.)* Observe that the match for this pair is 68.65%, which is not too much lower than 71.31% obtained by manual sentiment labeling. This shows

Data Pair	Match %
a. Ar(Manl.Sent) - Ar(Auto.Sent)	63.89
b. Ar(Manl.Sent) - En(Manl.Trans., Manl.Sent)	71.31
c. Ar(Manl.Sent) - En(Manl.Trans., Auto.Sent)	68.65
d. Ar(Manl.Sent) - En(Auto.Trans., Manl.Sent)	57.21
e. Ar(Manl.Sent) - En(Auto.Trans., Auto.Sent)	62.49
f. En(Manl.Trans., Manl.Sent) - En(Auto.Trans., Manl.Sent)	60.08
g. En(Manl.Trans., Manl.Sent) - En(Manl.Trans., Auto.Sent)	66.51
h. En(Auto.Trans., Manl.Sent) - En(Auto.Trans., Auto.Sent)	69.58

Table 4: Match percentage between pairs of sentiment labelled BBN datasets.

Data Pair	Match %
a. Ar(Manl.Sent) - Ar(Auto.Sent)	78.65
b. Ar(Manl.Sent) - En(Auto.Trans., Manl.Sent)	71.05
c. Ar(Manl.Sent) - En(Auto.Trans.-Auto.Sent)	78.11
d. En(Auto.Trans, Manl.Sent) - En(Auto.Trans., Auto.Sent)	78.80

Table 5: Match percentage between pairs of sentiment labelled Syria datasets.

that the English sentiment system is performing rather well. (One would not expect it to get a match greater than 71.31%.) More importantly, the English sentiment system shows a competitive result of 62.49% when run on the automatically translated text (row e.), which makes this choice a viable option for sentiment analysis of non-English texts. This result is inline with previous findings in Information Retrieval (Nie et al., 1999) and Text Classification (Amini and Goutte, 2010).

Rows d. and e. compare *Ar(Manl.Sent.)* with manual and automatic sentiment labeling of automatic translations. Since automatic translation from Arabic to English is fairly difficult, we expect these match percentages to be lower than those in rows b. and c., and that is exactly what we observe. However, it is unexpected to find the number for row e. to be higher than that of row d. We find the same pattern for corresponding data pairs in the Syrian tweets as well (rows b. and c. in Table 6). This suggests that certain attributes of automatically translated text mislead humans with regards to the true sentiment of the source text. However, these same attributes do not seem to affect the automatic sentiment analysis system as much. Since the NRC sentiment analysis system is largely reliant on word-sentiment associations and does not use syntax-based features, it is possible that syntactic abnormalities introduced by automatic translation impact human perception of sentiment. However, this supposition needs to be validated by future work.

Row f. shows that manual and automatic translation lead to only about 60% match in manually annotated sentiment labels with each other. Row g. shows accuracy of the English automatic sentiment analysis system on the manually translated text (assuming the English sentiment labels as gold). The result of 66.51% is very close to human agreement on manually translated data (68%), which demonstrates the high quality of the English sentiment analysis system. Row h. shows accuracy of the English automatic sentiment analysis system on the automatically translated text (assuming the English sentiment labels as gold). In this case, the system's accuracy of 69.58% is higher than the human agreement on automatically translated text (65.7%), which again shows that automatic translation greatly impacts sentiment perceived by humans.

We manually examined several tweets from the BBN dataset to understand why humans incorrectly annotate a tweet's automatic translation. Most of the cases were due to bad translation where sentiment words either disappeared or were replaced with words of opposite sentiment. In some cases, the translation was affected by typos on the Arabic side. Table 6 shows some examples. Often the mistranslations occurred due to word sense ambiguity. For example, عقارب has two meanings: *scorpions* and *clock arms*. In example 1 (metaphorically stating that relatives can hurt like scorpion bites), the word is mistranslated, leading to neutral (instead of negative) sentiment.

1. Bad auto. translation: mistranslation of ambiguous words		
Post	الدنيا علمتني ان اكثر الاقارب عقارب	negative
Auto.Trans.	the minimum taught me that more relatives clock	neutral
Manl.Trans.	Life has taught me that most of the relatives are scorpions	negative
2. Bad auto. translation: mistranslation of ambiguous words		
Post	ليتنى اعيش في مكان لا تنقطع عنه الثلوج	positive
Auto.Trans.	i wish i live in a place not cut off by snow	negative
Manl.Trans.	I wish I live in a place where snow never stops falling	positive
3. Bad auto. translation: sarcasm is hard to translate		
Post	لسه الخير لقدام تسرب المي موجودة من زمان	negative
Auto.Trans.	you're still good in front of the leakage of water existed from time	positive
Manl.Trans.	Expect more good to come, water has been leaking since a long time	negative

Table 6: Examples where the automatic translation was annotated a sentiment different from the sentiment of the original Arabic tweet, but whose original sentiment was correctly predicted by the English sentiment system. The manual translations are also listed for reference.

One reason why the automatic sentiment analysis system correctly annotates several automatically translated instances (where manual annotations of the translation may fail), is that the system can learn an appropriate model even from mistranslated text — especially when automatic translation makes consistent errors. For example, اللهم انصر (*Oh God grant victory to*) has been consistently translated to *God forsake*. All tweets having this phrase are correctly annotated as positive by our system, but were marked negative by the human annotators.

Caveats: The automatic systems employed in these experiments, i.e., Arabic sentiment analysis, English sentiment analysis, and SMT systems, exhibit state-of-the-art performance; nevertheless, further improvements are possible. The Arabic sentiment system will benefit from extended sentiment lexicons and features derived specifically for the Arabic language. The English sentiment analysis system can be further adapted to the peculiarities of machine-translated texts, which are notably different from regular English. The current translation system has been trained on non-tweet data that results in a high percentage of out-of-vocabulary words on our datasets. In our experiments, we assumed that all texts are written in Levantine dialect of the Arabic language. However, tweets can have a mixture of dialects or even a mixture of languages (e.g., Arabic and English). Addressing these factors will give even more insight on how sentiment is altered on translation, in specific contexts.

9 Conclusions

We presented a set of experiments to systematically study the impact of English translation (manual and automatic) on sentiment analysis of Arabic social media posts. Our experiments show that automatic sentiment analysis of English translations (even of automatic translations) can lead to competitive results—results that are similar to that obtained by current state-of-the-art Arabic sentiment analysis systems. Our results also show that automatic sentiment analysis of automatic translations outperforms the manual sentiment annotations of the automatically translated text. This suggests that SMT errors impact human perception of sentiment markedly more than automatic sentiment systems. This is an interesting avenue for future exploration. We also show that translated texts tend to lose some of the sentiment information and there is a relatively higher percentage of neutral instances in the translated text than in the original dataset. The resources created as part of this project (Arabic sentiment lexicons, Arabic sentiment annotations of social media posts, and English sentiment annotations of their translations) are made freely available.[5]

Acknowledgments

Thanks to Kareem Darwish and Eshrag Refaee for sharing their data. We thank Colin Cherry, Samuel Larkin, and Marine Carpuat for helpful discussions.

[5]http://www.purl.com/net/ArabicSentiment

References

Muhammad Abdul-Mageed, Mona Diab, and Sandra Kübler. 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. *Computer Speech & Language*, 28(1):20 – 37.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Portland, Oregon.

Khurshid Ahmad, David Cheng, and Yousif Almas. 2006. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*.

Mohammed Al-Kabi, Amal Gigieh, Izzat Alsmadi, Heider Wahsheh, and Mohamad Haidar. 2013. An opinion analysis tool for colloquial and standard Arabic. In *Proceedings of the 4th International Conference on Information and Communication Systems*, ICICS '13.

Massih-Reza Amini and Cyril Goutte. 2010. A co-classification approach to learning from multilingual corpora. *Machine learning*, 79(1-2):105–121.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 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. Association for Computational Linguistics.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Jerome Bellegarda. 2010. Emotion analysis using latent affective folding and embedding. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 1–9, Los Angeles, California.

Anthony C. Boucouvalas. 2002. Real time text-to-emotion engine for expressive Internet communication. *Emerging Communication: Studies on New Technologies and Practices in Communication*, 5:305–318.

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

Boxing Chen and Xiaodan Zhu. 2014. Bilingual sentiment consistency for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 607–615, Gothenburg, Sweden, April. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 224–233. Association for Computational Linguistics.

Samhaa R El-Beltagy and Ahmed Ali. 2013. Open issues in the sentiment analysis of Arabic social media: A case study. In *Proceedings of the 9th International Conference on Innovations in Information Technology*, pages 215–220. IEEE.

Ahmed El Kholy and Nizar Habash. 2012. Orthographic and morphological processing for English—Arabic statistical machine translation. *Machine Translation*, 26(1-2):25–45, March.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '11, pages 42–47.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Nizar Habash, Owen Rambow, and Ryan Roth. 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, April. The MEDAR Consortium.

Minqing Hu and Bing Liu. 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. ACM.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pages 125–132, New York, NY. ACM.

Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, L Alfonso Ureñalópez, and A Rturo Montejoráez.

2012. Sentiment analysis in Twitter. *Natural Language Engineering*, pages 1–28.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, page 976.

Saif M. Mohammad and Peter D. Turney. 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.

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

Saif M. Mohammad and Tony (Wenda) Yang. 2011. Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the ACL Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, pages 70–79, Portland, OR, USA.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 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.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, *SEM '12, pages 246–255, Montréal, Canada.

Ahmed Mourad and Kareem Darwish. 2013. Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA '13, pages 55–64.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17:95–135, 1.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–81. ACM.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, LREC '10, pages 1320–1326, Valletta, Malta, May.

Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 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.

Eshrag Refaee and Verena Rieser. 2014a. An Arabic Twitter corpus for subjectivity and sentiment analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC '14, Reykjavik, Iceland, May. European Language Resources Association.

Eshrag Refaee and Verena Rieser. 2014b. Subjectivity and sentiment analysis of Arabic Twitter feeds with limited resources. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, page 16.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 73–80, Dublin, Ireland, August.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 553–561, Stroudsburg, PA, USA. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 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.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA, June.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 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. Association for Computational Linguistics.