Journal of Artificial Intelligence Research 55 (2016) 95-130 Submitted 03/15; published 01/16

How Translation Alters Sentiment

Saif M. Mohammad SAIF.MOHAMMAD@NRC-CNRC.GC.CA
National Research Council Canada

Mohammad Salameh MSALAMEH@UALBERTA.CA
University of Alberta

Svetlana Kiritchenko SVETLANA.KIRITCHENKO@NRC-CNRC.GC.CA
National Research Council Canada

Abstract

Sentiment analysis research has predominantly been on English texts. Thus there exist many sentiment resources for English, but less so for other languages. Approaches to improve sentiment analysis in a resource-poor focus language include: (a) translate the focus language text into a resource-rich language such as English, and apply a powerful English sentiment analysis system on the text, and (b) translate resources such as sentiment labeled corpora and sentiment lexicons from English into the focus language, and use them as additional resources in the focus-language sentiment analysis system. In this paper we systematically examine both options. We use Arabic social media posts as stand-in for the focus language text. We show that sentiment analysis of English translations of Arabic texts produces competitive results, w.r.t. Arabic sentiment analysis. We show that Arabic sentiment analysis systems benefit from the use of automatically translated English sentiment lexicons. We also conduct manual annotation studies to examine why the sentiment of a translation is different from the sentiment of the source word or text. This is especially relevant for building better automatic translation systems. In the process, we create a state-of-the-art Arabic sentiment analysis system, a new dialectal Arabic sentiment lexicon, and the first Arabic–English parallel corpus that is independently annotated for sentiment by Arabic and English speakers.

1. Introduction

The term *sentiment analysis* is most commonly used to refer to the goal of determining the valence or polarity of a piece of text, whether it is positive, negative, or neutral. However, it can more generally refer to determining one's attitude towards a particular target or topic. Automatic sentiment analysis of text, especially social media posts, has a number of applications in commerce, public health, and public policy development. In the past two decades, a vast majority of research 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. With improvements in statistical machine translation systems over the last decade, we no longer have to rely on strictly monolingual sentiment analysis systems—at least two other alternatives may be viable:

(a) Run an English sentiment analysis system, using English resources, on English translations of the focus language text.

(b) Use a focus-language sentiment analysis system that employs focus-language resources and translations of English resources into the focus language.

In this paper we systematically examine both options. We use Arabic social media posts as a specific instance of focus 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 and English-to-Arabic translation systems. We outline the advantages and disadvantages of each of the methods listed above, and conduct quantitative and qualitative experiments to determine impact of translation on sentiment. As benchmarks we use manually determined sentiment labels of the Arabic posts.

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 text automatically translated from another language into their native language?
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?
6. Can Arabic posts sentiment analysis systems benefit from additional training data that is an automatic translation of sentiment-labeled English tweets or from additional sentiment lexicons that are automatic translations of existing English lexicons?
7. Do automatic translations of words have the same sentiment associations as the original source words (as listed in the source language lexicons, say)? And if not, what are the different reasons that lead to discrepancies?

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 sentiment. 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.

We show that while it is difficult to obtain improvement in an Arabic sentiment analysis systems simply by adding more training data that is a translation of existing labeled English corpus, these systems benefit from the use of automatically translated English sentiment lexicons. By examining a subset of translated lexicon entries we show that close to 90% of the entries are valid even in the focus language. A word and its automatic translation may not convey the same sentiment because of poor translation quality or because the word and its translation are used differently in the two languages.

In the process of developing these experiments to study how translation impacts sentiment, we created a new dialectal Arabic sentiment lexicon and a state-of-the-art Arabic sentiment analysis system by porting NRC-Canada's competition winning system (Mohammad, Kiritchenko, & Zhu, 2013; Kiritchenko, Zhu, & Mohammad, 2014b) to Arabic. We also created a substantial amount of sentiment labeled data pertaining to Arabic social media texts and their English translations. 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. All of these sentiment lexicons and sentiment-labeled corpora are made freely available.[1]

We begin with a survey of related work in sentiment analysis of English, sentiment analysis in Arabic, and work in cross-lingual sentiment analysis (Section 2). In Section 3 we present our core method to systematically study the impact of translation on sentiment. In Section 4 we describe how we developed the components needed for our experiments: translations of Arabic texts into English, translations of English resources into Arabic, sentiment-labeled data in Arabic and English, an English sentiment analysis system, and an Arabic sentiment analysis system. In Section 5, we present results of the experiments on translating focus language text into English for application of an English sentiment analysis system. We also conduct qualitative and quantitative studies to investigate some of the reasons why sentiment is impacted on translation. For example, we find that sentiment expressions are often mistranslated into neutral expressions, however automatic sentiment analysis systems are able to recover to some extent from these errors. In Section 6, we present results of the experiments on translating English resources into Arabic and using features drawn from them in Arabic sentiment analysis. We also describe a manual annotation study on the extent to which automatic Arabic translations have the same sentiment as the source English words. Finally, we present conclusions and future directions in Section 7.[2]

2. Related Work

Over the last decade, there has been an explosion of work exploring various aspects of sentiment analysis in English texts: detecting subjective and objective sentences; classifying sentences as positive, negative, or neutral; detecting the person expressing the sentiment and the target of the sentiment; and applying sentiment analysis in health, commerce, and disaster management. Surveys by Pang and Lee (2008), Liu and Zhang (2012), and Mohammad (2016) give details of many of these approaches. However, there is less work on Arabic texts. In the sub-sections below, we briefly outline relevant sentiment analysis research on English texts, on Arabic texts, and on texts in one language using resources from another (multilingual sentiment analysis).

[1] http://www.purl.org/net/ArabicSA

[2] Some early findings of this work were first presented in Salameh, Mohammad, and Kiritchenko (2015).

2.1 Sentiment Analysis of English Social Media

English sentiment analysis systems have been applied to many different kinds of texts including customer reviews, newspaper headlines (Bellegarda, 2010), novels (Boucouvalas, 2002; John, Boucouvalas, & Xu, 2006; Mohammad & Yang, 2011), emails (Liu, Lieberman, & Selker, 2003; Mohammad & Yang, 2011), blogs (Neviarouskaya, Prendinger, & Ishizuka, 2011; Genereux & Evans, 2006; Mihalcea & Liu, 2006), 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, Bhayani, and Huang (2009), Pak and Paroubek (2010), Agarwal, Xie, Vovsha, Rambow, and Passonneau (2011), Thelwall, Buckley, and Paltoglou (2011), Brody and Diakopoulos (2011), Aisopos, Papadakis, Tserpes, and Varvarigou (2012), Bakliwal, Arora, Madhappan, Kapre, Singh, and Varma (2012). A survey by Martínez-Cámara, Martín-Valdivia, Ureñalópez, and Montejoráez (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, Kozareva, Nakov, Rosenthal, Stoyanov, & Ritter, 2013; Rosenthal, Ritter, Nakov, & Stoyanov, 2014; Pontiki, Galanis, Pavlopoulos, Papageorgiou, Androutsopoulos, & Manandhar, 2014). The NRC-Canada system (Kiritchenko et al., 2014b) ranked first in these competitions, and we use it in our experiments. Notably, the system makes extensive use of sentiment lexicons and handles negation appropriately.[3] We summarize that system in Section 4.3.

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. Negation in MSA is expressed through negation particles, but in some dialects (Egyptian) it is expressed using a circumfix.

There have been a few studies tackling sentiment analysis of Arabic texts (Ahmad, Cheng, & Almas, 2006; Farra, Challita, Assi, & Hajj, 2010; Abdul-Mageed, Diab, & Korayem, 2011; Badaro, Baly, Hajj, Habash, & El-Hajj, 2014). There is also a shared task on detecting sentiment intensity of Arabic phrases (Kiritchenko, Mohammad, & Salameh, 2016).[4] The works most closely related to ours are the studies of sentiment analysis of Arabic social media (Al-Kabi, Gigieh, Alsmadi, Wahsheh, & Haidar, 2013; Ahmed, Pasquier, & Qadah, 2013; El-Beltagy & Ali, 2013; Mourad & Darwish, 2013; Abdul-Mageed, Diab, & Kübler, 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

[3] Zhu, Guo, Mohammad, and Kiritchenko (2014a) show that the impact of negation cannot be properly captured by simply reversing the polarity of its scope.

[4] http://alt.qcri.org/semeval2016/task7/

separates subjective from objective sentences and then classifies the subjective into positive or negative instances. The authors compiled several datasets from multiple social media resources that included chatroom messages, tweets, forum posts, and Wikipedia Talk pages. The datasets were manually labeled by two native Arabic speakers. However, these resources have not been made publicly available yet. Abdul-Mageed and Diab (2014) also used data from several resources to compile and build SANA, a large-scale, multi-genre, multidialect lexical resource. SANA covers Egyptian and Levantine dialects as well as MSA. Abbasi, Chen, and Salem (2008) deployed Arabic morphological, syntactic and stylistic features for sentiment analysis of Arabic web forums. For efficient feature selection, they adopted an Entropy Weighted Genetic Algorithm (EWGA). 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 4.4.2.

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. We test our system on that dataset. However, the dataset they provided is a superset of the data they had originally used in their experiments (Refaee & Rieser, 2014a). Thus, the performances of automatic sentiment analysis systems applied 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, Banea, and Wiebe (2007) used English resources to automatically generate a Romanian subjectivity lexicon using an English–Romanian dictionary. The generated lexicon was then used to classify Romanian text. Balahur and Turchi (2014) conducted a study to assess the performance of statistical sentiment analysis techniques on machine-translated texts. Opinion-bearing English phrases from the New York Times Text (2002–2005) corpus were split into training and test datasets. An English sentiment analysis system was trained on the training dataset and its prediction accuracy on the test set was found to be about 68%. Next, the training and test datasets were automatically translated into German, Spanish, and French using publicly available machine-translation engines (Google, Bing, and Moses). The translated test sets were then manually corrected for errors. Then for German, Spanish, and French, a sentiment analysis system was trained on the translated training set for that language and tested on the translated-and-corrected test set. The authors observe that these German, Spanish, and French sentiment analysis systems obtain accuracies in the low sixties (and thus not very much lower than 68%). Contrary to this work, our study uses original text from the focus language, its manual and automatic translations, as well as both manual and automatic sentiment assignments to systematically examine the effect of translation on sentiment. Further, we use several external sentiment resources as well as their translations within state-of-the-art sentiment systems. Also, German, Spanish, and French are much closer to English, than Dialectal Arabic is to English. Finally, we deal with noisy social media texts as opposed to more polished news media texts. There also exists research on using sentiment analysis to improve machine translation, such as the work by Chen and Zhu (2014), but that is beyond the scope of this paper.

3. Method for Determining the Impact of Translation on Sentiment

To systematically study the impact of translation on sentiment analysis, we propose two experimental setups corresponding to (a) and (b) described in the Introduction:

- Setup A: Translate Arabic text into English (manually and automatically) and annotate the English text for sentiment (manually and automatically). Compare the sentiment labels assigned to the translated English text with manual sentiment annotations of the Arabic text. The more similar the sentiment annotations are, the less is the impact of translation.

- Setup B: Translate sentiment annotated corpora and lexicons from English into Arabic (automatically), and use them as additional resources in supervised Arabic sentiment classification. Compare the sentiment labels assigned by this system with manual sentiment annotations of the Arabic text. The more similar the sentiment annotations are, the less is the impact of translation.

3.1 Impact of Translation on Sentiment - Setup A: Translating the Focus Language Text to English

With Setup A we explore how translation of text from Arabic to English impacts its sentiment. Specifically, we analyze the performance of an English sentiment analysis system, using English resources, on automatic translations of Arabic social media texts. The setup is outlined below:

- Identify or compile an Arabic social media dataset. We will refer to it as *Ar*. [*Ar* comes from the first two letters 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.)*.

- 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.)*. This acts as a baseline system.

- 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.

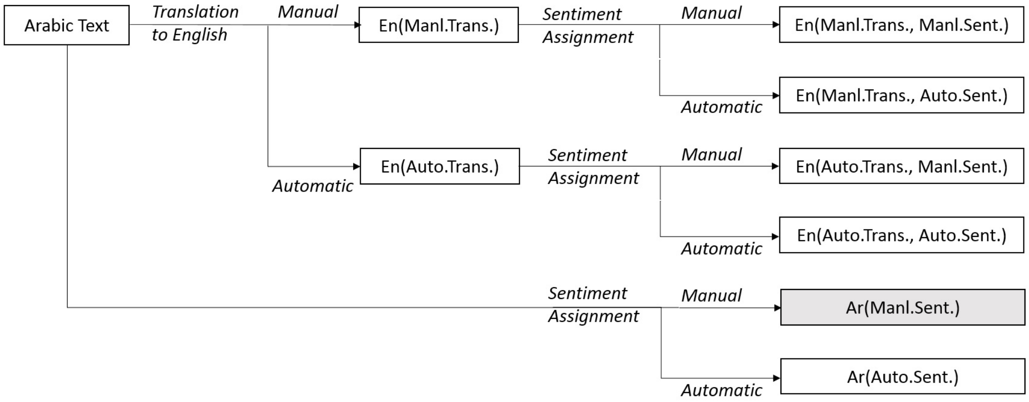

Figure 1: Setup A: Translating the focus language text to English. We compare sentiment labels between Ar(Manl.Sent.) (shown in a shaded box) and other datasets shown on the right side of the figure. Ar(Manl.Sent.) is the original Arabic text manually annotated for sentiment.

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 manually translated from Arabic to English. The comparison will also show, for example, whether positive tweets tend to be translated into neutral tweets, and to what extent. Furthermore, the results will demonstrate 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 5, we provide an analysis of several such comparisons for two different Arabic social media datasets.

DATA and RESOURCES: The list of all corpora and lexicons used in Setup A is shown in Table 1. 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, Malchiodi, Devlin, Stallard, Matsoukas, Schwartz, Makhoul, Zaidan, & Callison-Burch, 2012).[5] 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.

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 *Syria dataset*.[6] Note, however, that manual translations of the Syrian tweets are not

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

[6] The number of instances chosen to be in the BBN and Syria datasets is somewhat arbitrary; however, we were constrained by the funds available for manual sentiment annotations on these datasets and their translations.

Resource	**Number of instances**			
	positive	**negative**	**neutral**	**total**
a. Focus language (Arabic) corpora				
(and their English translations)				
BBN posts	498	575	126	1,199
Syrian tweets	448	1,350	202	2,000
b. Resources explored by the baseline Arabic sentiment system				
Automatic lexicons:				
Arabic Emoticon Lexicon	22,962	20,342	-	43,304
Arabic Hashtag Lexicon	13,118	8,846	-	21,964
Arabic Hashtag Lexicon (dialectal)	11,941	8,179	-	20,128
c. Resources explored by the English sentiment system				
Manual lexicons:				
Bing Liu's Lexicon	2,006	4,783	-	6,789
MPQA Subjectivity Lexicon	2,718	4,911	570	8,199
NRC Emotion Lexicon	2,317	3,338	8,527	14,182
Automatic lexicons:				
NRC Emoticon Lexicon	38,312	24,156	-	62,468
NRC Hashtag Lexicon	32,048	22,081	-	54,129

Table 1: Resources used in Setup A. (Note 1: The focus language corpora are split into test and training folds as part of cross-validation experiments. Note 2: 'NRC Emotion Lexicon' and 'NRC Emoticon Lexicon' are very similar in spelling, but they are two different lexicons.)

available. In our automatic sentiment analysis experiments, the focus language corpora (BBN dataset and Syria dataset) are each split into test and training folds as part of cross-validation experiments.

We use a number of manually and automatically created English sentiment lexicons in our English sentiment analysis system (as shown in row c. of Table 1). We compare the accuracies obtained by the English sentiment analysis system with an Arabic sentiment analysis system, for which we create new Arabic word–sentiment association lexicons as described in Section 4.4.1. These lexicons are called the Arabic Hashtag Lexicon, the Dialectal Arabic Hashtag Lexicon, and the Arabic Emoticon Lexicon.

3.2 Impact of Translation on Sentiment - Setup B: Translating English Sentiment Resources to the Focus Language

With Setup B we explore how translation of text from English into Arabic impacts its sentiment. Specifically, we analyze the change in performance of an Arabic sentiment analysis system when it is allowed to also make use of automatic translations of English sentiment lexicons and corpora. The setup is outlined below:

- Identify an Arabic social media dataset. Manually annotate it for sentiment and split the corpus into test and training subsets. We will refer to the test corpus as *Ar* and the sentiment-labeled test corpus as *Ar(Manl.Sent.)*.

- Identify or create suitable Arabic sentiment lexicons.
- Identify suitable English sentiment lexicon(s) and a corpus of English tweets labeled for sentiment.
- Automatically translate the English corpus and lexicon into Arabic. We will refer to the Arabic translation of the corpus as *Ar(Auto.Trans.)* and Arabic translation of the lexicon as *ArLex(Auto.Trans.)* [*Auto.* is for automatic; *Lex* is for lexicon.]
- Train separate Arabic sentiment analysis systems using each of the following sets of resources:
  1. the Arabic training corpus only;
  2. the Arabic training corpus and the Arabic translation of the English corpus;
  3. the Arabic training corpus and the Arabic sentiment lexicon;
  4. the Arabic training corpus, the Arabic sentiment lexicon, and the Arabic translation of the English lexicon.

  Apply each of the Arabic sentiment analysis systems on the test set *Ar*.

Figure 2 depicts this setup. Once the various sentiment-labeled datasets are created, we can compare the automatically labeled sets with the manual sentiment annotations of *Ar(Manl.Sent.)*, and calculate accuracies of the automatic labeling. These accuracies will help answer questions such as: how useful automatically translated English sentiment resources are for Arabic sentiment analysis. We also perform a manual annotation study on a subset of the automatically translated resources to determine different kinds of errors that result from the automatic translations. In Section 6, we provide an analysis of these experiments on different English resources.

DATA and RESOURCES: The list of all corpora and lexicons used in Setup B is shown in Table 2. We chose the Arabic portion of the BBN Arabic-Dialect–English Parallel Text as the primary Arabic social media dataset for Setup B. Specifically, we use the same subset of 1200 Levantine dialectal sentences, which we refer to as the *BBN posts* or *BBN dataset*. As the English corpus, we choose the SemEval-2013 Task 2 (Sentiment Analysis in Twitter) training dataset (Wilson et al., 2013) for our experiments because just as the BBN dataset, this is a dataset of social media posts. Further, it is already manually annotated for sentiment.

There are several sentiment lexicons for English. We chose four manually created lexicons for our experiments: NRC Emotion Lexicon (Mohammad & Turney, 2010; Mohammad & Yang, 2011), Bing Liu Lexicon (Hu & Liu, 2004), MPQA Subjectivity Lexicon (Wilson, Wiebe, & Hoffmann, 2005), and AFINN (Nielsen, 2011). We also experiment with the lexicons automatically generated from tweets by Kiritchenko et al. (2014b): NRC Emoticon Lexicon (a.k.a. Sentiment140 lexicon) and NRC Hashtag Sentiment Lexicon. These lexicons helped obtain the best results in sentiment analysis shared task competitions (Mohammad et al., 2013; Kiritchenko, Zhu, Cherry, & Mohammad, 2014a; Zhu, Kiritchenko, & Mohammad, 2014b).

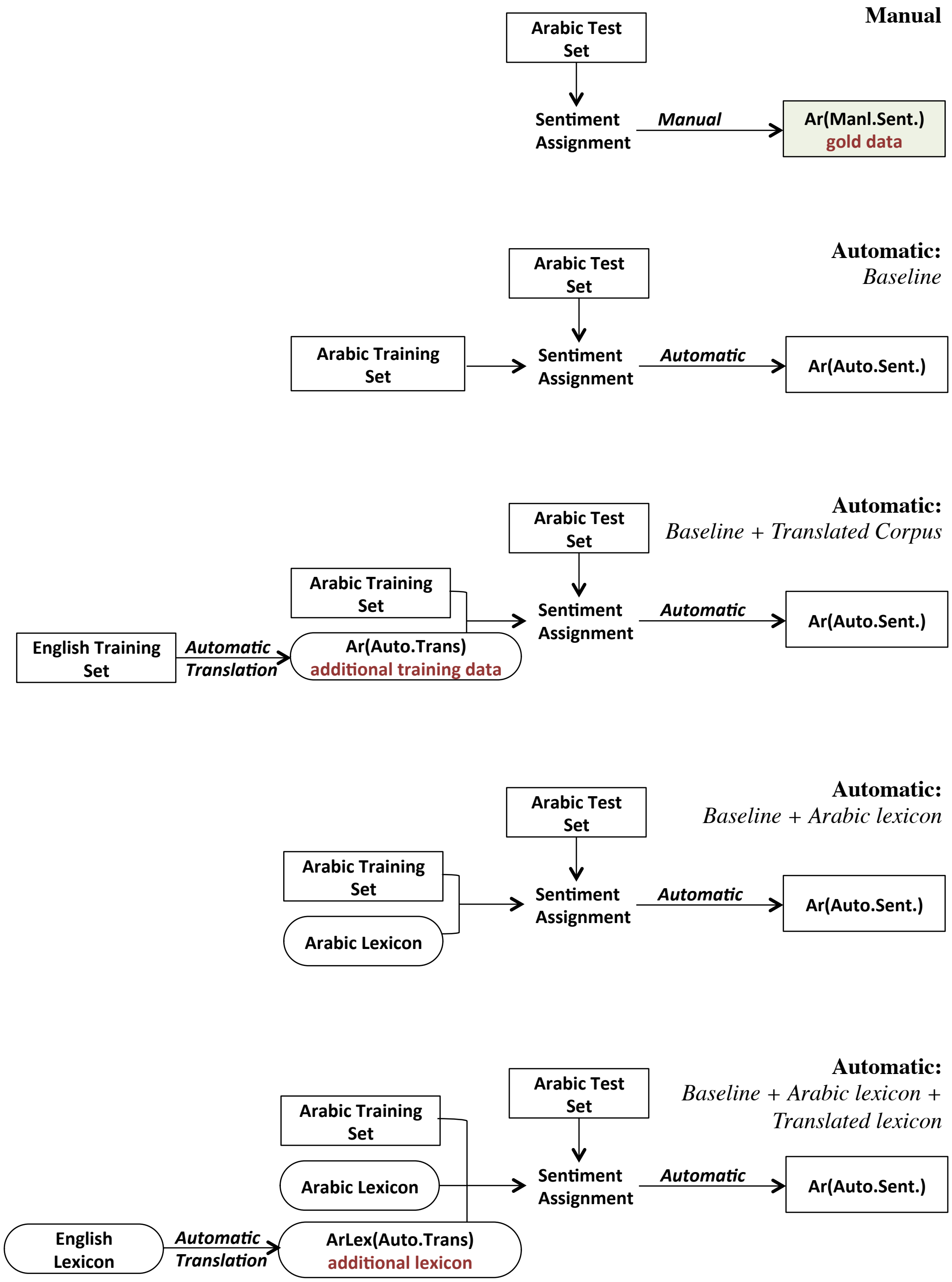

Figure 2: Setup B: Translating English sentiment resources to the focus language. We compare sentiment labels between Ar(Manl.Sent.) (shown in a shaded box) and other datasets shown on the right side of the figure. Ar(Manl.Sent.) is the original Arabic text manually annotated for sentiment.

Resource	**Number of instances**			
	positive	**negative**	**neutral**	**total**
a. Focus language (Arabic) corpora				
BBN posts	498	575	126	1,199
b. Resources used by the Arabic sentiment system				
Automatic lexicons:				
Arabic Hashtag Lexicon (dialectal)	11,941	8,179	-	20,128
c. English resources translated into Arabic				
Sentiment-labeled corpus:				
SemEval-2013 Task 2 corpus	3,620	1,549	4,743	9,912
Manual lexicons:				
AFINN	878	1,598	-	2,476
Bing Liu's Lexicon	2,006	4,783	-	6,789
MPQA Subjectivity Lexicon	2,718	4,911	570	8,199
NRC Emotion Lexicon	2,317	3,338	8,527	14,182
Automatic lexicons:				
NRC Emoticon Lexicon	15,210	11,530	-	26,740
NRC Hashtag Lexicon	18,341	14,241	-	32,582

Table 2: Resources used in Setup B. (Note 1: In our automatic sentiment analysis experiments, focus language corpora are split into test and training folds as part of cross-validation experiments. Note 2: Automatic translations of the English resources into Arabic were done using Google Translate. Some entries, especially in the automatic lexicons, were left untranslated because Google Translate had no information on them.)

4. Capabilities Needed for Performing the Experiments

The experimental setups described above involve several component tasks: generating translations manually and automatically (Section 4.1), manually annotating Arabic and English texts for sentiment (Section 4.2), automatic sentiment analysis of English texts (Section 4.3), and automatic sentiment analysis of Arabic texts (Section 4.4). We describe each of them in the sub-sections below.

4.1 Generating Translations

Setup A requires certain Arabic corpora translated into English, whereas Setup B requires some English resources (corpus and lexicon) to be translated into Arabic. The two subsections below describe how we obtained these translations.

4.1.1 Generating English Translations

The BBN dialectal Arabic dataset comes with manual translations into English. We generate automatic English translations of the BBN posts and the Syrian tweets by employing our in-house multi-stack phrase-based machine translation (MT) system, Portage (Cherry & Foster, 2012). This statistical machine translation (SMT) system is trained on data from

OpenMT 2012. We preprocess the training data by segmenting the Arabic source side of the training data with MADA 3.2 (Habash, Rambow, & Roth, 2009), using Penn Arabic Treebank (PATB) segmentation scheme as recommended by El Kholy and Habash (2012). Since the different forms of Arabic characters Alif (ا آ أ إ) and Ya (ي ى) are used interchangeably, we normalize these characters to a bare Alif ا and dotless Ya ى, respectively. This normalization decreases the sparcity of Arabic tokens and improves translation. The English side of the training data is lower-cased 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, Marton, & Resnik, 2008; Cherry & Foster, 2012). A KN-smoothed 5-gram language model is trained on the English Gigaword and the target side of the parallel data.

4.1.2 Generating Arabic Translations

For Setup B, we run the SemEval-2013 English tweets dataset (Wilson et al., 2013) through Google Translate to obtain Arabic translations.[7] Even though Google Translate is a phrase-based statistical MT 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. Thus we also use Google Translate to translate into Arabic the words in each of the English sentiment lexicons listed in Table 2. Note that Google Translate is unable to translate some words in these lexicons. Table 2 gives the number of words translated as well as a break down by sentiment category (positive, negative, and neutral). All of the translated lexicons are made freely available.[8] We do not generate manual translations of these lexicons, but in Section 6.1, we describe a study where the automatic translations are examined by an Arabic speaker.

4.2 Creating Sentiment Labeled Data in Arabic and English

Manual sentiment annotations were performed on the crowdsourcing platform CrowdFlower[9] for three BBN datasets and two Syria datasets:

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

The Questionnaire for 3. is shown below. The questionnaire for 2. is very similar, except that it states that the text was created by manual translation of Arabic posts. The questionnaire for 1. is also very similar to 3., except that it is in Arabic and it states that the

[7] Since our in-house system, Portage, is designed to translate text from Arabic to English, but not the other way round, we use the publicly available Google Translate for the experiments in Setup B. Google Translate: https://translate.google.com

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

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

target texts are posts from social media (no mention of translations in this questionnaire).

Questionnaire for 3: Judge the sentiment of the posts

General Instructions:
- Attempt HITs only if you are a native speaker of English.
- Your responses are confidential.
- It is possible that the occasional post may have a swear word or express something offensive. The text is no different than something one might find in any public forum.

Task-Specific Instructions:
You will be given English sentences that were translated from Arabic using an automatic machine translation system. The translations may be ungrammatical and hard to understand. If the translation system was unable to translate a word, it shows that word with the UNK symbol, representing unknown.
Select the option that best captures the sentiment being conveyed in the sentence:
- positive
- negative
- neutral
- uncertain OR both positive and negative

Select “positive” if the sentence shows a positive attitude (possibly toward an object or event). For example:
- I hope every year you will be in good shape
- To be honest I don’t know what to say in this story, the nicest sensation

Select “negative” if the sentence shows a negative attitude (possibly toward an object or event). For example:
- The new Spiderman movie is terrible
- This government will make us bankrupt

Select “neutral” if the sentence shows a neutral attitude (possibly toward an object or event). For example:
- Add spices, onion and sauce
- This expresses truly our relation with Israel

Select “Uncertain OR both positive and negative” if the sentence shows an uncertain attitude OR if the sentence expresses both positive and negative attitude. For example:
- The strange that the forward glass of car is not broken yet
- I like ice cream but hate chocolate chips on it

Actual HIT

The sentence below was translated into English from Arabic by a computer algorithm. The sentence may be ungrammatical and hard to follow. Additionally, the system was unable to translate some Arabic words. These words are shown with the “UNK” symbol.

Sentence: Especially companies to acknowledge will be a soft target UNK penetrate serwer commercial network .

Select the option that best captures the sentiment being conveyed in the sentence:
- positive
- negative
- neutral
- uncertain OR both positive and negative

Each post was annotated by at least ten annotators and the majority sentiment label was chosen. A very small number of instances were annotated with the label "uncertain OR both positive and negative". These instances were set aside and not included in further analysis. Table 3 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 original 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. 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 3. We use this agreement score as benchmark to compare performance of automatic sentiment systems (described below).

4.3 English Sentiment Analysis

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

A linear-kernel Support Vector Machine (Chang & 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 & Turney, 2010, 2013), Bing Liu Lexicon (Hu & Liu, 2004), and MPQA Subjectivity Lexicon (Wilson et al., 2005).

The NRC Emotion Lexicon has sentiment and emotion labels for about 14,000 words (Mohammad & Turney, 2010; Mohammad & Yang, 2011). These labels were compiled through Mechanical Turk annotations.[10] The Bing Liu Lexicon has about 6,800 words with sentiment labels (Hu & Liu, 2004). The lexicon was originally used for detecting sentiment of customer reviews. The MPQA Subjectivity Lexicon, which draws from the General Inquirer and other sources, has sentiment labels for about 8,000 words (Wilson et al., 2005).

[10] https://www.mturk.com/mturk/welcome

	positive	**negative**	**neutral**	**agreement**
BBN dataset				
a. Ar(Manl.Sent)	41.50	47.92	10.58	73.82
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 dataset				
d. Ar(Manl.Sent)	22.40	67.50	10.10	79.05
e. En(Auto.Trans., Manl.Sent)	14.25	66.15	19.60	76.10

Table 3: Class distribution (in percentage) of the sentiment annotated datasets.

We also used the automatically generated, tweet-specific lexicons NRC Hashtag Sentiment Lexicon and NRC Emoticon Lexicon (Kiritchenko et al., 2014b).[11] The sub-section below gives more details about how these lexicons were generated.

4.3.1 Generating English Sentiment Lexicons

The ablation experiments in the study by Mohammad et al. (2013) showed that the NRC-Canada sentiment analysis system benefited most from the use of the NRC Hashtag Sentiment Lexicon and the NRC Emoticon Lexicon. The NRC Hashtag Sentiment Lexicon was created as follows. A list of 77 seed words, which are synonyms of *positive* and *negative*, was compiled from the Rogets Thesaurus. Then, Twitter API was polled to collect tweets that had these words as hashtags. Not all tweets that have a positive hashtag or emoticon express positive sentiment. And similarly not all tweets that have a negative hashtag or emoticon express negative sentiment. Hashtags and emoticons can be used in tweets in complex ways, for example in sarcastic tweets. Nonetheless, a majority of the tweets with a positive hashtag or emoticon have been shown to be positive (and similarly for negative hashtags and emoticons). Thus the algorithm to extract positive and negative terms from tweets considers a tweet to be positive if it has a positive hashtag and negative if it 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 or negative category:

$$SenScore\,(w) = PMI(w, pos) - PMI(w, neg) \tag{1}$$

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) implies that the word tends to co-occur more with positive seeds than with negative seeds. Thus, it is likely to be associated with positive sentiment. Similarly, a negative score suggests that the word tends to co-occur more with negative seeds than with positive seeds. Thus, it is likely to be associated with negative sentiment. The magnitude of $SenScore\,(w)$ indicates the strength of the association.

The NRC Emoticon Lexicon (aka Sentiment140 Lexicon) was created in a similar fashion using emoticons ':)' and ':(' as seeds.

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

4.3.2 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, Schneider, O'Connor, Das, Mills, Eisenstein, Heilman, Yogatama, Flanigan, & Smith, 2011). Then, each tweet is represented as a feature vector.

The features:

- Word and character ngrams;
- POS: the number of occurrences of each part-of-speech tag;
- Negation: the number of 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 sentence and present in a lexicon, its sentiment score $score(w)$ is used to compute:
  - the number of tokens with $score(w) \neq 0$;
  - the total score $= \sum_{w \in tweet} score(w)$;
  - the maximal score $= max_{w \in tweet} score(w)$; and
  - the score of the last token in the tweet.

  These features are calculated for each lexicon separately.
- Manually created sentiment lexicons: For each of the three manual sentiment lexicons, the following features are computed:
  - the sum of positive scores for tweet tokens;
  - the sum of negative scores for tweet tokens.
- Style features: the presence/absence of all-cap words, hashtags, punctuation marks, emoticons, and elongated words.

4.4 Arabic Sentiment Analysis

We build an Arabic sentiment analysis system by reconstructing the NRC-Canada English system to deal with Arabic text. It extracts the same feature set as described in Section 4.3.2. We have also generated word-sentiment association lexicons using the process described in Section 4.3.1, but for Arabic words from Arabic tweets (more details in subsection 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.

4.4.1 Generating 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 different Arabic sentiment lexicons:

Lexicon	# seeds		# tweets		# entries in lexicon		
	pos	neg	pos	neg	unigram	bigram	trigram
Arabic Emoticon Lexicon	12	11	520,000	455,282	43,309	229,747	325,366
Arabic Hashtag Lexicon	109	121	209,784	37,209	22,007	128,814	233,481
Arabic Hashtag Lexicon (dialectal)	135	348	177,556	34,705	20,128	93,613	159,986

Table 4: Details of the Arabic Hashtag Lexicon, the Arabic Emoticon Lexicon, and the Dialectal Arabic Hashtag Lexicon.

- *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 (same as described in Section 4.3.1, and proposed first in Mohammad et al., 2013):

$$SentimentScore(w) = PMI(w, pos) - PMI(w, neg) \tag{2}$$

  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. 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 that included these seed words as hashtags. After filtering out duplicate tweets and retweets, we ended up with 209,784 positive unique tweets and 37,209 negative unique tweets. For each unigram, bigram, and trigram, w, that occurred at least five times in these tweets, *SenScore*(w) was calculated just as described in Section 4.3.1. We will refer to this lexicon as the *Arabic Hashtag Lexicon*.

- *Arabic Hashtag Lexicon (Dialectal)*: Refaee and Rieser (2014a) 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)*.

The number of seeds and tweets used to create the Arabic Hashtag Lexicon, the Arabic Emoticon Lexicon, and the Dialectal Arabic Hashtag Lexicon are shown in Table 4. The table also shows the number of unigram, bigram, and trigram entries in each of the lexicons.

Dataset	**BBN posts**	**Syrian tweets**
Sentiment classes	pos, neg, neu	pos, neg, neu
Number of instances	1199	2000
Most frequent class baseline	47.95	67.50
Human agreement benchmark	73.82	79.05
Our system, using all non-lexicon features and		
a. the Arabic Emoticon Lexicon features	62.40	78.35
b. the Arabic Hashtag Lexicon features	62.97	78.96
c. the Dialectal Arabic Hashtag Lexicon features	**65.31**	**79.35**
d. lexicon features from a., b., and c.	63.47	79.00

Table 5: Accuracy obtained using features from different automatically generated Arabic sentiment lexicons. The highest scores are shown in bold.

Arabic Sentiment Labeled Dataset	**MD**	**RR**	**BBN posts**	**Syrian tweets**
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	65.31	79.35

Table 6: Accuracy (in percentage) of sentiment analysis (SA) systems on various Arabic social media datasets.

4.4.2 Evaluation

Table 5 shows ten-fold cross-validation accuracies obtained on the BBN and Syria datasets using the various Arabic sentiment lexicons discussed above. Observe that the best results are obtained when using the Dialectal Arabic Hashtag Lexicon. Both, the Arabic Hashtag Lexicon and the Arabic Emoticon Lexicon features, when added to the Dialectal Arabic Hashtag Lexicon features did not result in an improvement in classification accuracy. Henceforth in the paper, we use the Dialectal Arabic Hashtag Lexicon as the only Arabic sentiment lexicon.

Existing sentiment-labeled Arabic datasets include the one described in Mourad and Darwish (2013), which we will refer to as *MD*, and the one described in Refaee and Rieser (2014a), which we will refer to as *RR*. We tested the Arabic sentiment system on MD and RR, as well as the two newly sentiment-annotated Arabic datasets—BBN posts and Syrian tweets. Table 6 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 datasets, 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 the previously published results on the MD dataset. The only previously published results on the RR dataset are on a small

Dataset	**BBN posts**	**Syrian tweets**
Sentiment classes	pos, neg, neu	pos, neg, neu
Number of instances	1199	2000
Most frequent class baseline	47.95	67.50
Human agreement benchmark	73.82	79.05
Our System		
a. All Features	65.31	79.35
b. All - lexicon features	61.98	79.35
c. All - ngram features	63.07	66.45
c1. All - word ngram features	64.72	77.71
c2. All - char. ngram features	63.31	78.40
d. All - style features	65.23	79.20
e. All - ngram features - style features	62.23	67.35
f. All - lexicon features - style features	61.90	79.45

Table 7: Ablation experiments showing accuracy on the BBN and Syria datasets. The larger the drop in performance when removing a feature set, the more useful that feature set is in classification.

subset (about 1000 instances) for which Refaee and Rieser (2014a) obtained an accuracy of about 87%. The results in Table 6 are for a larger dataset and so not directly comparable.

We determine the impact of different feature sets on performance by conducting ablation experiments, where we remove one set of features at a time and observe the change in performance. The larger the drop in accuracy, the more useful the removed feature set. Table 7 shows the ablation results on the BBN and Syria datasets.

Observe that for the BBN dataset the largest drop in performance occurs when we remove the sentiment lexicon features. This shows that the method for producing the sentiment lexicon was effective in generating useful word–sentiment association entries. Ngrams too are helpful in sentiment classification, especially for the Syria dataset. Removing the style features (features based on hashtags, exclamations, etc.) does not result in a large drop in performance for both datasets, and it is likely that the character ngram features subsume much of their discerning power. Row e. shows performance when we use only the sentiment lexicon features (no ngram features and no style features) and row f. shows performance when we use only the ngram features (no lexicon features and no style features). Observe that the performance using the lexicon features alone is rather competitive in the BBN dataset, suggesting that the automatically generated sentiment lexicons are able to capture term–sentiment association to a similar extent to what the supervised algorithm can learn from ngram features in training data. On the Syria dataset, ngrams alone produce results reaching human agreement levels. We believe this may be because of the markedly lower type to token ratio (lexical diversity) in the Syrian tweets and due to the skew in the dataset towards the negative class. (Table 15 in Section 5.2 shows type to token ratios in various datasets.)

	pos	neg	neu
BBN dataset			
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 dataset			
d. Ar(Auto.Sent)	20.60	75.30	4.10
e. En(Auto.Trans., Auto.Sent)	24.75	69.75	5.50

Table 8: Class distribution (in percentage) resulting from automatic sentiment analysis.

Data Pair	Match %
a. Ar(Manl.Sent) - Ar(Auto.Sent)	65.31
b. Ar(Manl.Sent) - En(Manl.Trans., Manl.Sent)	71.31
c. Ar(Manl.Sent) - En(Manl.Trans., Auto.Sent)	67.73
d. Ar(Manl.Sent) - En(Auto.Trans., Manl.Sent)	57.21
e. Ar(Manl.Sent) - En(Auto.Trans., Auto.Sent)	62.08
f. En(Manl.Trans., Manl.Sent) - En(Auto.Trans., Manl.Sent)	60.08
g. En(Manl.Trans., Manl.Sent) - En(Manl.Trans., Auto.Sent)	63.11
h. En(Auto.Trans., Manl.Sent) - En(Auto.Trans., Auto.Sent)	69.58

Table 9: Setup A: Match percentage between pairs of sentiment labeled BBN datasets.

5. Experiments on Sentiment after Translation - Setup A: Translating Focus Language Text into English

With Setup A (as described in Section 3.1) we analyze performance of an English sentiment analysis system, using English resources, on automatic translations of Arabic social media texts. Using the methods and systems described in Sections 4.1, 4.2, 4.3, and 4.4, we generated all the translations and the manually and automatically sentiment labeled datasets mentioned in Section 3.1's Experimental Setup (also shown in Figure 1). Table 8 shows the distribution of positive, negative, and neutral classes in various datasets that have been automatically labeled with sentiment. These percentages can be compared with those in Table 3 (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 9 and 10 show how similar the sentiment labels are across various pairs of datasets for the BBN posts and the Syrian tweets, respectively. For example, row a. in Table 9 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 65.31%

Data Pair	Match %
a. Ar(Manl.Sent) - Ar(Auto.Sent)	79.35
b. Ar(Manl.Sent) - En(Auto.Trans., Manl.Sent)	71.05
c. Ar(Manl.Sent) - En(Auto.Trans., Auto.Sent)	79.16
d. En(Auto.Trans, Manl.Sent) - En(Auto.Trans., Auto.Sent)	76.80

Table 10: Setup A: Match percentage between pairs of sentiment labeled Syria datasets.

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 73.82%. 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 67.73%, which is not too much lower than 71.31% obtained by manual sentiment labeling. This shows 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 62.08% 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 cross-lingual information retrieval (Nie, Simard, Isabelle, & Durand, 1999) and text classification (Amini & 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 10). 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. We conduct experiments to explore the reasons behind this in Section 5.1.

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 the accuracy of the English automatic sentiment analysis system on the manually translated text (assuming the English sentiment labels as gold). 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.

5.1 Qualitative Analysis of Why Translations Differ in Sentiment from the Source Text

As can be seen from the results of the experiments in the previous section, translations of text often do not preserve the original sentiment. Further, there exist a number of instances where the manual sentiment annotations of automatic translations differ from the sentiments of the original Arabic text, but the automatic English sentiment analysis system correctly predicts the sentiment of the original Arabic text. We now describe a study we conducted to determine why that is — what some of the main reasons are, and how frequently these reasons come into play.

We started by creating a dataset where manual annotations of the Arabic texts disagreed with manual annotations of the translations. Specifically, from the BBN dataset, we created instances composed of:

a. the original Arabic tweet,
b. manually determined sentiment of the Arabic tweet (positive, negative, or neutral),
c. manual English translation of the Arabic tweet,
d. manually determined sentiment of the translation (positive, negative, or neutral).

We kept only those instances where b. differed from d. We further filtered this set keeping only those instances where the automatic English sentiment analysis system correctly predicted b. These instances were arranged in decreasing order of inter-annotator agreement of sentiment annotation on the Arabic texts. Since the annotation task is time intensive, we wanted to annotate those instances where there is high confidence in the sentiments of the original Arabic texts. The top 100 instances were presented to a judge who spoke both English and Arabic fluently. We will refer to this dataset as the *BBN Manl.Trans. Disagreement Pairs*. For each of these 100 instances, the judge was asked why in their opinion b. and d. differ. The precise directions are as shown below:

Annotation Guidelines

For each instance (row), tell us why you think the manually annotated sentiment of the English translation differs from the original sentiment of the Arabic post.

Codes:

1. Bad translation
   a. sentiment words disappear
   b. sentiment words added
   c. sentiment words replaced with opposite sentiment words
   d. something other than sentiment words has (also) caused disagreement (may be ill-formed text, may be grammar, may be the position of negators like not and never, or tense, or auxiliaries, etc.)
2. Translation is reasonable (sentiment-wise), but the same sentence can be viewed as having one sentiment in the Arabic speaking population and different sentiment in the English-speaking population due to cultural and life-style differences.

3. Do not know.

Note:

- Some of the codes have sub-categories. So you can enter 1b, 1c, etc. You can even enter just 1 if none of the sub-categories apply.
- As you annotate, if you discover new categories, you can add them to the list of codes, and use the new codes as well.
- If more than one code applies to an instance, separate them by a comma. For example, you can say "1a, 1d" or "1b, 1c, 2".

We then repeated the annotation procedure, but now for instances involving *automatically* translated texts:

a. the original Arabic tweet,
b. manually determined sentiment of the Arabic tweet (positive, negative, or neutral),
c. *automatic* English translation of the Arabic tweet,
d. manually determined sentiment of the translation (positive, negative, or neutral).

Just as before, we kept only those instances where b. differed from d., and only those instances where the automatic English sentiment analysis system correctly predicted b. The 100 instances with highest inter-annotator agreement on Arabic sentiment annotation were presented to the judge. We will refer to this dataset as the *BBN Auto.Trans. Disagreement Pairs*. The judge was asked why in their opinion b. and d. differ. Since automatic translations exist for both the BBN and the Syria datasets, this annotation was done for 100 instances from the Syria dataset as well. We will refer to this dataset as the *Syria Auto.Trans. Disagreement Pairs.*

5.1.1 Annotation Results

It took a human judge 12 hours to annotate the three sets (100 instances each) described above. The distribution of the reasons for disagreement between the sentiment of the original text and the sentiment of its translation in 100 instances from each dataset is shown in Table 11. Since the total number of instances in this study is 100, the number for each reason (code) is also the corresponding percentage. Note that since an instance can be annotated to belong to more than one reason, the percentages do not sum to 100%. Also, since the annotator could choose a broad reason category code (for example, 1.), if none of its sub-categories apply (for example, 1a. or 1b.), the sum of entries for the subcategories need not be equal to the number of entries in the subsuming reason category.

The judge marked only a handful of instances in the "do not know" category and did not add any new reason categories. This shows that the judge was largely able to determine the reason for disagreement between the two manual sentiment annotations involved (one of the original Arabic post and one of the English translation), in terms of the other reasons pre-specified.

	Percentage of Disagreement Pairs		
	BBN dataset		**Syria dataset**
Reason for Disagreement	**Manl.Trans.**	**Auto.Trans.**	**Auto.Trans.**
1. Bad Translations	41	95	91
1a. sentiment words disappear	11	58	80
1b. sentiment words added	0	1	1
1c. sentiment words replaced with words of opposite sentiment	9	37	8
1d. sentiment changed due to ill-formed text, grammar, etc	26	13	12
2. Cultural differences	73	10	26
3. Do not know	2	3	4

Table 11: Class distribution of the reasons for disagreement between the sentiment of the original text and the sentiment of its translation in 100 instances from three datasets. Note that the entries represent both the number and percentage of instances, since each subset has 100 instances in all.

A large percentage of instances in the manually translated disagreement sets were affected by what the judge thought were cultural differences. However, bad translation was still a significant cause of disagreement. Our own cursory examination of the BBN dataset also gave us the impression that the manual translations could have been better.

Nonetheless, in contrast to the manually translated disagreement sets, the automatically translated disagreement sets had a markedly high proportion of instances (more than 90%) where the bad translation led directly to the disagreement in sentiment. More specifically, automatic translations seem to often mistranslate sentiment expressions such that either they do not appear in the translation or they appear as neutral terms in the translation (58% of instances in BBN Auto.Trans. and 80% in Syria Auto.Trans.).

Very few instances pertaining to 1b. were found in the data. This is not surprising since one would not expect the translator to add sentiment where there is none.

5.1.2 Discussion

Table 12 shows examples for the disagreement categories resulting from the *manual* translation of the BBN subset. (We do not show examples of 1b. because of lack of data for this sub-category.) For each of the sub-categories present, the table shows the original Arabic post, the BBN-provided manual translation, and the comments from the judge. Table 13 shows examples for each of the disagreement categories resulting from the *automatic* translation of the BBN subset.

Discussions with the judge revealed that the following phenomenon commonly led to 1a., 1c., 1d., and 2.:

- Ambiguous words: Often a word with many meanings, where one sense has a certain sentiment (positive, negative, or neutral) and another sense has a different sentiment, can be mistranslated into the wrong sense leading to sentiment disagreement. This is more common in automatically translated text, but occurs sometimes even in manually

1a. Sentiment words disappear		
Post	**جبهم برص** ،،، شو اعملّهم يعني ؟	negative
Manl.Trans.	So what, what can I do to them	neutral
Comments	The bolded text is an Arabic expression that literally means “let them be loved by a Gecko”. It expresses disgust or anger. This part was left untranslated by the human translator.	
1c. Sentiment words replaced with words of opposite sentiment		
Post	لك تاني شغله هو لسا بالانعاش	negative
Manl.Trans.	secondly, is he still in the **refreshment** room?	neutral
Comments	An ambiguous Arabic word was mistranslated into *refreshment* instead of *recovery room*. It is surprising that the human translator made this mistake.	
1d. Sentiment changed due to ill-formed text, grammar, tense, etc		
Post	لسه الخير لقدام تسرب المي موجودة من أيام انجاز المشروع	negative
Manl.Trans.	The good is still coming, the water leak is from the day they had the project	neutral
Comments	The post is supposed to shows sarcasm by saying “expect more good to come”, meaning “the worse is yet to come”. This expression is widely used in Arabic conversations to mean something negative. This is also an example of 2.	
2. Cultural Differences		
Post	بصراحة .. لا تعليق لدي ..	negative
Manl.Trans.	honestly... I have no comment...	neutral
Comments	Although the post does not seem to be literally negative, but in many Arabic conversations it is used to express a negative opinion—similar to “I am speechless”.	
Post	مع اني ما لقيت الهلال من بيتنا ؟	negative
Manl.Trans.	although I didn’t see the crescent from home	neutral
Comments	The post was associate with negative sentiment as observing the crescent moon in Islam is associated with the beginning of a month or a holiday	

Table 12: Examples of different reasons for disagreement between the sentiment of the original text and the sentiment of its manual translation.

translated text—especially if the translation is done by crowdsourcing, and quality control checks are not stringent. Even human translators have on occasion been tempted to use Google Translate.

- Sarcasm: Sarcasm can sometimes be hard to detect, even for humans, and even when detected, upon translation, differences in cultural and language norms can mean that the translation no longer appears sarcastic. See example for 1d. in Table 12.
- Metaphors: Metaphors, such as the example for 1c. in Table 13, are also hard to translate—again more so by the automatic system, but to some extent even by humans. These have often been translated into neutral or opposite sentiment expressions, contributing to 1a. or 1c.
- Word-reordering: Automatic translations can often lead to poor word-reordering in the target language, and this has sentiment implications when the original post has negation terms. Missing or misplaced negation term can lead to a different sentiment. Sarcasm is also greatly affected by word-reordering. See example 1d. in Table 13.

Most current statistical machine translation systems are evaluated using an ngram-based evaluation metric (BLEU). However, the metric often misses (or does not penalize enough)

1a. Sentiment words disappear		
Post	مبارة **نااايمة** صراحة	negative
Auto.Trans.	match UNK frankly .	neutral
Comments	The bolded word is dialectal Arabic typo not translated by the system. It is meant to be “sleeeeeping” i.e., the match was boring	
1c. Sentiment words replaced with words of opposite sentiment		
Post	الدنيا علمتني ان اكثر الاقارب عقارب	negative
Auto.Trans.	the minimum taught me that more relatives clock	neutral
Comments	عقارب has two meanings: *scorpions* and *clock arms*. Also الدنيا means either “word” or “lower”. The post is supposed to metaphorically state that “the world has taught me relatives can hurt like scorpion bites”. The post is mistranslated, leading to neutral (instead of negative) sentiment.	
1d. Sentiment changed due to ill-formed text, grammar, tense, etc		
Post	وما بتعرف شو يعني نظافة	negative
Auto.Trans.	and you know what , i mean , the cleanliness	positive
Comments	The correct translation is “she does not know what cleanliness means”. Word reordering and missing a negation led to text with seemingly positive sentiment.	
2. Cultural Differences		
Post	وانتا مش عارف اصلا كيف عاملينها وشو حاطين فيها ..	negative
Auto.Trans.	and you don’t know how they are doing and what they are putting in place .	neutral
Comments	The post is perceived by Arab annotators as being said in a conversation to express negative attitude toward an object	
Post	اللهم ارحم موتانا اللهم اشف مرضانا	positive
Auto.Trans.	God have mercy on our dead God cure our patients	negative
Comments	Supplications in Arabic is annotated as positive, although it contains lots of negative phrases. The tweet is annotated as positive with confidence of 1.0, and its automatic translation as negative with confidence of 0.7, thus showing cultural differences in perceiving this tweet	

Table 13: Examples of different reasons for disagreement between the sentiment of the original text and the sentiment of its automatic translation.

mistranslations caused by many of the phenomenon listed above. Thus, as is evident here, this means that automatically generated translations can often carry misleading sentiment.

5.2 Quantitative Analysis of Features of Translations that Impact Sentiment

In the previous section, we qualitatively analyzed why human annotation of sentiment on translations is difficult. In this section we quantitatively explore:

I what causes automatic translations to be inferior to manual translations in terms of preserving sentiment. (Recall from Table 9 that b. and c. have higher scores than d. and e.)

II why automatic translation, more error prone as it may be, offers some advantages to an automatic sentiment analysis system as compared to human annotations. (Recall from Table 9 that row d. has a lower score that row e.)

Although it is hard to prove causation, we hope the experiments below shed more light into the features of translated texts that impact sentiment.

Dataset	BBN Manl.Trans.	BBN Auto.Trans.
Our System		
a. All features	67.73	62.08
b. All - lexicon features	63.73	60.74

Table 14: Accuracy of the sentiment analysis system on the manual and automatic translations, with and without sentiment lexicons.

Dataset	#types	#tokens	#types/#tokens
BBN posts (Arabic)	6,054	11,928	0.5075
BBN posts (Manl.Trans.)	3,592	16,609	0.2163
BBN posts (Auto.Trans.)	3,108	16,660	0.1866
Syrian tweets (Arabic)	11,667	35,983	0.3242
Syrian tweets (Auto.Trans.)	6,731	57,153	0.1178

Table 15: Lexical diversity in the datasets.

The qualitative analysis in the previous section suggests that one of the main reasons for I may be the fact that sentiment words in the original text are translated to more neutral terms. We test this quantitatively through ablation experiments on the English translations (manual and automatic), by observing the effect of removing sentiment lexicon features. Table 14 shows the results. Observe that sentiment lexicon features are more helpful in the manual translations (improve results by 4 percentage points) than in the automatic translations (improves results by only 1.34 points).

Our hypothesis for why the automatic sentiment analysis system correctly annotates several automatically translated instances where manual annotations of the translation may fail (II), is that the sentiment 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.

If this were true, then we surmise that automatic translations will have a lower lexical diversity than manual translations. That is, automatic translations have a lower word type (unique term) to token ratio than manual translations. Table 15 shows the number of types and tokens in the original Arabic BBN and Syria datasets and also their translations. For automatic translations we removed all UNK tokens before determining these counts.

First, we note that even human translations have a lower type to token ratio than the original source text. Additionally, observe that as hypothesized, the type to token ratio is markedly higher in manual translations as compared to automatic translations of the same text. This supports the hypothesis that the SMT system translates source tokens more consistently. Since the automatic sentiment analysis system is trained on these consistently translated text with the original sentiment labels of the source text, it is still able to determine the true sentiment. However, since human sentiment annotators see many instances where the sentiment terms are mistranslated into neutral terms, they are unable to determine the true sentiment.

System	**Accuracy (in percentage)**
a. Baseline (uses word ngram and style features from training fold)	61.98
b. Baseline + Arabic translation of English corpus	
English corpus: SemEval task 2 training corpus	42.78
c. Baseline + Arabic translation of English lexicon	
i. English lexicon: AFINN	63.41
ii. English lexicon: Bing Liu Lexicon	62.99
iii. English lexicon: MPQA	61.91
iv. English lexicon: NRC Emotion Lexicon	**63.48**
v. English lexicon: NRC Emoticon Lexicon	62.40
vi. English lexicon: NRC Hashtag Lexicon	61.73
d. Baseline + Arabic lexicon	
i. Arabic lexicon: Arabic Emoticon Lexicon	62.40
ii. Arabic lexicon: Arabic Hashtag Lexicon	62.97
iii. Arabic lexicon: Arabic Hashtag Lexicon (dialectal)	**65.31**
e. Baseline + Arabic lexicon + Arabic translation of English lexicon	
Arabic lexicon: Arabic Hashtag Lexicon (dialectal)	
i. English lexicon: AFINN	65.73
ii. English lexicon: Bing Liu Lexicon	66.15
iii. English lexicon: MPQA	65.15
iv. English lexicon: NRC Emotion Lexicon	**66.23**
v. English lexicon: NRC Emoticon Lexicon	66.15
vi. English lexicon: NRC Hashtag Lexicon	64.22
f. Baseline + Arabic Hashtag Lexicon (dialectal)	
+ Arabic translation of NRC Emotion Lexicon	**66.57**

Table 16: Setup B: Cross-validation experiments on the BBN dataset. The highest score overall and the highest scores in c., d., and e. are shown in bold.

6. Experiments on Sentiment after Translation - Setup B: Translating Sentiment Resources from English to the Focus Language

We now describe sentiment analysis experiments where we automatically translate resources from English to the focus language (Arabic) to improve accuracy of a sentiment analysis system operating on texts in the focus language (Setup B). We implemented Setup B as described in Section 3.2 and Figure 2, and using the capabilities described in Section 4. The translations were obtained using Google Translate. The Arabic portion of the BBN dataset was used as the primary focus language text. Our baseline system performs ten-fold cross-validation on this dataset using word ngrams and style features described in Section 4.4. All other systems use additional resources—some originally created from Arabic sources and some that are translations of English resources.

Table 16 shows the accuracy of these automatic sentiment analysis systems. (Generated sentiment labels are compared to manual annotation by Arabic speakers.) Comparing rows a. and b. we can infer that simply translating sentiment-labeled tweets from English into Arabic and using them as additional training data for Arabic sentiment analysis leads to poor results. As we saw in Section 5, the language produced by a machine translation system differs substantially from the corresponding natural language. Therefore, a sentiment analysis system cannot fully benefit from an additional labeled machine-translated corpus when asked to annotate natural language text at test time. Similar observations were reported in work by Balahur and Turchi (2014). It is possible that better results may be obtained by using one of the many domain-adaptation techniques proposed in the literature, however, we leave that for future work.[12]

We also conducted experiments by adding the Arabic translations of the English lexicons to the baseline system of row a.—see results in rows c.i. through c.vi. of the table. The best results are obtained by using Arabic translations of the NRC Emotion Lexicon (row c.iv.).

Row d. shows results obtained when using the baseline system with additional features from an Arabic sentiment lexicon. As shown in Section 4.4, the Dialectal Arabic Hashtag Lexicon outperformed other Arabic lexicons—those results are shown again here for completeness and convenience. We compare accuracies obtained by rows e. and f. with d. to determine if using additional features from Arabic translations of English sentiment lexicons is beneficial. Observe that most of the translated English lexicons help obtain higher accuracies than that of row d.iii. (65.31%). The best results obtained with translations of an English sentiment lexicon are from using the NRC Emotion Lexicon. Using the NRC Emotion Lexicon along with the Dialectal Arabic Hashtag Lexicon in addition to the baseline system (row f.) gives a slight further improvement (66.57%).

Thus in the sentiment analysis of Arabic social media posts, it is difficult to extract benefit from automatic translations of English sentiment labeled sentences. However, improvements can be obtained using automatic translations of English sentiment lexicons.

6.1 Manual Examination of Automatically Translated Entries from a Sentiment Lexicon

As shown above, lexicons created by translating existing ones in other languages can be beneficial for automatic sentiment analysis, even if one has good lexicons in the focus language (such as the Dialectal Arabic Hashtag Lexicon for Arabic). 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 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 the focus language 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, if required. The error categories 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.

[12] Sampling the English corpus to obtain a similar class distribution as in the Arabic dataset led to only small improvements.

	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 17: 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 18: Percentage of erroneous entries in the translated NRC Emotion Lexicon that are assigned to each error category.

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 led to the best results in our experiments (Table 16). 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 17 shows the results of the human annotation study. Of the 100 positive entries examined, 85 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, 11.67% of the translated entries were deemed incorrect for Arabic.

Table 18 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%).

7. Conclusions

Much of the work in sentiment analysis is focused on English texts. Thus, most other languages have limited sentiment resources. In this paper we conducted several experiments exploring two broad approaches for improving sentiment analysis in Arabic social media text with the help of English resources and state-of-the-art translation systems: (a) translate the focus language text into a resource-rich language such as English, and apply a powerful English sentiment analysis system on the translation, and (b) translate resources such as sentiment labeled corpora and sentiment lexicons from English into the focus language, and use them as additional resources in the focus-language sentiment analysis system. Our goal was to systematically study the impact of translation (manual and automatic) on sentiment.

Our experiments show that automatic sentiment analysis of English translations (even of automatic translations) of Arabic texts can lead to competitive results—results that are similar to that obtained by current state-of-the-art Arabic sentiment analysis systems. Similar findings have been reported for other tasks, such that Information Retrieval (Nie et al., 1999) and Text Classification (Amini & Goutte, 2010). Surprisingly, 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. Furthermore, we conduct qualitative and quantitative studies to investigate why we observe these results. We find that sentiment expressions are often mistranslated into neutral expressions when translated. Additionally, automatic translation often makes consistent errors in translating terms, and since the automatic system learns term–sentiment associations from training data, it can learn that the mistranslated word is a cue for the true sentiment, thus recovering from the error. Sarcasm, metaphoric expressions, and incorrect word-reordering are some other common reasons why translations fail to preserve sentiment. Finally, we observe that even correctly translated texts are sometimes marked as having a different sentiment than what speakers of the source language believe it to be. Thus, sentiment, at least to some extent is dependent on the cultural context of the annotator.

We also conducted experiments on translating English resources into Arabic to help improve Arabic sentiment analysis systems. Specifically, we found that using automatic Arabic translations of many freely available English sentiment lexicons improved accuracy. However, experiments that simply added translated, sentiment-labeled, English tweets data to existing Arabic training data resulted in a drop in accuracy. On manual examination of a subset of the automatic translations of the English lexicon entries, a native speaker of Arabic marked close to 90% as appropriate (that is, the Arabic word had the same sentiment association as its English source word). The annotator, who is fluent in English as well, categorized the remaining 10% of the entries into different error classes—reasons because of which the entries were not valid in Arabic. Mistranslation, cultural differences, and different sense distributions in Arabic and English, were the primary reasons for errors in the automatic translation of entries in the sentiment lexicon.

Caveats: The automatic systems employed in these experiments, i.e., Arabic sentiment analysis, English sentiment analysis, and machine translation, exhibit state-of-the-art performance; nevertheless, further improvements are possible. The Arabic sentiment analysis system will possibly benefit from 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 machine translation system has been trained on non-tweet data that results in a high percentage of out-of-vocabulary words on our datasets. Tweets can have a mixture of dialects or even a mixture of languages (e.g., Arabic and English). Addressing these factors in future work will give even more insight into how sentiment is altered on translation, in specific contexts.

Data: All of 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.[13]

Acknowledgments

Thanks to Kareem Darwish and Eshrag Refaee for sharing their data. Thanks to Norah Alkharashi for annotating translations with reasons for why sentiment was not preserved. We thank Colin Cherry, Samuel Larkin, and Marine Carpuat for helpful discussions.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, *26*(3), 12:1–12:34.

Abdul-Mageed, M., & Diab, M. (2014). SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, LREC '14. European Language Resources Association.

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

Abdul-Mageed, M., Diab, M. T., & Korayem, M. (2011). Subjectivity and sentiment analysis of modern standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 587–591.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pp. 30–38, Portland, Oregon.

Ahmad, K., Cheng, D., & Almas, Y. (2006). Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*.

Ahmed, S., Pasquier, M., & Qadah, G. (2013). Key issues in conducting sentiment analysis on Arabic social media text. In *Proceedings of the 9th International Conference on Innovations in Information Technology*, pp. 72–77. IEEE.

[13] http://www.purl.org/net/ArabicSA

Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012). Textual and contextual patterns for sentiment analysis over microblogs. In *Proceedings of the 21st International Conference on World Wide Web Companion*, WWW '12 Companion, pp. 453–454, New York, NY, USA.

Al-Kabi, M., Gigieh, A., Alsmadi, I., Wahsheh, H., & Haidar, M. (2013). An opinion analysis tool for colloquial and standard Arabic. In *Proceedings of the 4th International Conference on Information and Communication Systems*, ICICS '13.

Amini, M.-R., & Goutte, C. (2010). A co-classification approach to learning from multilingual corpora. *Machine learning*, *79*(1-2), 105–121.

Badaro, G., Baly, R., Hajj, H., Habash, N., & 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)*, pp. 165–173, Doha, Qatar.

Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., & Varma, V. (2012). Mining sentiments from tweets. In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pp. 11–18, Jeju, Republic of Korea.

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

Bellegarda, J. (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*, pp. 1–9, Los Angeles, California.

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

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 562–570, Stroudsburg, PA, USA.

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

Chen, B., & Zhu, X. (2014). Bilingual sentiment consistency for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 607–615, Gothenburg, Sweden. Association for Computational Linguistics.

Cherry, C., & Foster, G. (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*, pp. 427–436.

Chiang, D., Marton, Y., & Resnik, P. (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, pp. 224–233.

El-Beltagy, S. R., & Ali, A. (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*, pp. 215–220. IEEE.

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

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

Genereux, M., & Evans, R. P. (2006). Distinguishing affective states in weblogs. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 27–29, Stanford, California.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (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, pp. 42–47.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Tech. rep., Stanford University.

Habash, N., Rambow, O., & 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*, pp. 102–109, Cairo, Egypt. The MEDAR Consortium.

Hu, M., & 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, pp. 168–177, New York, NY, USA. ACM.

John, D., Boucouvalas, A. C., & Xu, Z. (2006). Representing emotional momentum within expressive Internet communication. In *Proceedings of the 24th International Conference on Internet and Multimedia Systems and Applications*, pp. 183–188, Anaheim, CA. ACTA Press.

Kiritchenko, S., Mohammad, S., & 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.

Kiritchenko, S., Zhu, X., Cherry, C., & 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)*, pp. 437–442, Dublin, Ireland.

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

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Aggarwal, C. C., & Zhai, C. (Eds.), *Mining Text Data*, pp. 415–463. Springer US.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, pp. 125–132, New York, NY. ACM.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureñalópez, L. A., & Montejoráez, A. R. (2012). Sentiment analysis in Twitter. *Natural Language Engineering*, 1–28.

Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, p. 976.

Mihalcea, R., & Liu, H. (2006). A corpus-based approach to finding happiness. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 139–144. AAAI Press.

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

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. *Emotion Measurement.*

Mohammad, S. M., Kiritchenko, S., & 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.

Mohammad, S. M., & 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*, pp. 26–34, LA, California.

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

Mohammad, S. M., & Yang, T. W. (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, pp. 70–79, Portland, OR, USA.

Mourad, A., & Darwish, K. (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, pp. 55–64.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Affect analysis model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17*, 95–135.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (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*, pp. 74–81. ACM.

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*, pp. 93–98.

Pak, A., & Paroubek, P. (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, pp. 1320–1326, Valletta, Malta.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

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

Refaee, E., & Rieser, V. (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. European Language Resources Association.

Refaee, E., & Rieser, V. (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*, p. 16.

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pp. 73–80, Dublin, Ireland.

Salameh, M., Mohammad, S., & 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*, pp. 767–777, Denver, Colorado.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Wilson, T., Kozareva, Z., Nakov, P., Rosenthal, S., Stoyanov, V., & 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.

Wilson, T., Wiebe, J., & 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, pp. 347–354, Stroudsburg, PA, USA.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., & 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*, pp. 49–59. Association for Computational Linguistics.

Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014a). An empirical study on the effect of negation words on sentiment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*, Vol. 14.

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