Automatic Machine Translation for Arabic Tweets

Fatma Mallek, Ngoc Tan Le and Fatiha Sadat

Abstract Twitter is a continuous and unlimited source of data in natural language, which is particularly unstructured, highly noisy and short, making it difficult to deal with traditional approaches to automatic Natural Language Processing (NLP). The current research focus on the implementation of a phrase-based statistical machine translation system for tweets, from a complex and a morphological rich language, Arabic, into English. The first challenge is prepossessing the highly noisy data collected from Twitter, for both the source and target languages. A special attention is given to the pre-processing of Arabic tweets. The second challenge is related to the lack of parallel corpora for Arabic-English tweets. Thus, an out-of-domain corpus was incorporated for training a translation model and an adaptation strategy of a bigger language model for English tweets was used in the training step. Our evaluations confirm that pre-processing tweets of the source and target languages improves the performance of the statistical machine translation system. In addition, using an in-domain data for the language model and the tuning set, showed a better performance of the statistical machine translation system from Arabic to English tweets. An improvement of 4 pt. BLEU was realized.

Keywords Microblogs · Twitter · Statistical machine translation (SMT)
Language model · Parallel corpus · Arabic · English

1 Introduction

Since decades, the Machine Translation (MT) has been the subject of interest of many researches in the domain of Natural Language Processing (NLP). The principal purpose of MT is to translate a natural language into another one. This task seems very simple to a human expert in translation, but not for the computer. For this reason, the research in the domain of MT is ongoing and many machine learning

F. Mallek (✉) · N. Tan Le · F. Sadat
Université du Québec À Montréal, Montreal, Canada
e-mail: mallek.fatma@courrier.uqam.ca

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_6

methods have been proposed recently. In the literature, we distinguish several types of approaches in MT (1) the example-based approach, (2) the rule-based approach, (3) the statistical approach and finally (4) the hybrid approach. The two first approaches are less and less used due to the considerable human efforts they need to build the rules and the dictionaries. For this reason, current research in the field of MT tends to choose automatic approaches based on machine learning, such as statistical ones.

Statistical Machine Translation (SMT) is one of the most popular approaches, especially the Phrase-based one (PBSMT). It reposes on monolingual and parallel corpora. Parallel corpora consists of a collection of bilingual texts which are generally aligned at the level of the sentences or paragraphs, i.e. texts in the source language with their translations in the target language.

Nowadays, with the emergence of social media sites and the fluidity of digital data, the MT applications are focused more and more on this kind of data [20, 46]. Especially in the recent years, with the political events around the world, it seems very interesting to understand what other people published via the social media in any language. Twitter, is considered as one of the famous social networking platform. It reaches the place of the second most popular social networking site in the world [14]. Registered users can post short messages or tweets, of up to 140 characters at a time. They can also follow messages of other users. According to the study of the company's own blog, the number of active users reached 310 millions per month. Also, about 500 millions of tweets are published by users everyday in more than 40 different languages.[1] In front of this diversity of languages, the company of Twitter has proposed to its users the option to translate their tweets. So, in January 2015, the Bing Translator was integrated in Twitter's platforms. Yet, this option do not reach a good translation performance.[2]

The fluidity and the linguistic characteristics of tweets published continuously has been the subject of several studies in NLP on many topics such as sentiment analysis [31, 38], event detection [8], named entities recognition [29] and machine translation [20, 21, 46]. Tweets are considered as very short texts and are written in non-standard format. Users of Twitter make many orthographic mistakes and often express their ideas in more than one language in the same time. For these reasons, translating the content of social media texts is considered as a very challenging task [9]. This task is more complex when a morphological rich language such as Arabic is involved [17].

Actually the linguistic resources for Arabic social media, such as parallel corpus (Arabic-other language or other language-Arabic) and the NLP tools are very limited. In [31, 38, 42] the authors collected parallel corpora for Arabic-English tweets, but these data collections are not enough to build an efficient machine translation for Arabic-English tweets. Also the data collected in [20] is not open source. For that, in our study, we will train the statistical translation model using a public parallel corpus (Modern Standard Arabic (MSA)/English) and then we will apply adapting strategies for the tweet's contents.

[1] https://about.twitter.com/fr/company.

[2] https://support.twitter.com/articles/20172133.

This research requires to tackle several challenges at the same time and raises several questions:

- What are the linguistic characteristics of the Arabic language used in the social media platforms? Is it closed to the Modern Standard Arabic?
- What are the pre-processing steps for the machine translation applications, when we deal with a morphological rich and complex languages such as Arabic?
- Is it a good choice to build a SMT system for tweets, using an out-of-domain corpus? What is the efficient strategy to adapt the MT system in order to translate noisy and short texts such as tweets?

2 Arabic Language Challenges Within NLP and Social Media

Arabic language is a semitic language that is spoken by more than 300 millions persons, all over the world [18]. There is 28 letters in the Arabic alphabet, and sentences are written from right to left. In the literature, we can distinguish more than one kind of Arabic language: Modern Standard Arabic (MSA) and dialectal Arabic (DA), or colloquial language [40]. MSA is the formal language used in educational and scripted speeches. On the other hand, DA is the daily language of several people in the Arab world that dominantly used on the social media websites [41]. However, the texts on the social media, especially in tweets, are mixed between MSA and DA and with many variations [40]. In the sections below, we detailed the challenges of the MSA language for MT and NLP applications. Afterwards, we focus on the characteristics of the language used on social media.

2.1 Modern Standard Arabic Challenges for MT

The orthographic, morphological and syntactic characteristics of the Arabic language raise many issues and challenges for Machine Translation (MT) applications.

Orthographic Ambiguity: Actually Arabic words can have different semantic and/or syntactic meanings depending on the marks that are added to the consonants. These marks are named Diacritics. For example, the diacritics added to the sentence “كتب الولد” handle to two different sentences with different meanings “كَتَبَ الوَلَدُ” [ktb Alwld] (the boy wrote) or “كُتُبُ الوَلَدِ” [kutubu Alwldi] (the book’s boy) [10]. Such semantic ambiguity in the orthographic representations of sentences deeply affects the quality of the translated Arabic texts [43].

Another orthographic challenge with Arabic is caused by the different spelling ways of some letters. The letter “Hamza” or “Alif” “أ” [A] are differently written.

Also, the letter “Ya” is sometimes written as a dotless “Ya” (‘(ى)’). These orthographic ambiguities, lead to poor probabilistic estimations of words in SMT applications [43].

2.1.1 Morphological Complexity

We distinguish two types of sentences in Arabic language: the nominal sentences and the verbal sentences. A nominal sentence is composed by a subject and an attribute, which may be a qualifying adjective, a complement adverb, a complement of an object, etc. For example, the sentence “الطقس جميل.” [AlTqs jmyl] (The weather is nice), is a nominal sentence that does not contains a verb (the verb to be (is) in English). It is composed by a subject and an adjective. In the verbal sentences, the verb is presented and it inflects subject (gender, number and person), aspect, voice, or mood.

Actually, Arabic words can belong to one of three categories: verbs, names and particles. Many words, semantically different, can be derived from the same root, by applying different patterns. Hence, the extraction of the lemma from an Arabic word is a challenging task and requires the use of a morphological analyzer [39]. Otherwise, Arabic words are agglutinated words, composed by an inflected word form (base) and attachable clitics. So, a word in Arabic can correspond to multiple English words. This problem complicates the alignment between an Arabic and an English sentence and increases the OOV (out-of-vocabulary) rate in a SMT system [39]. Table 1 shows the derived words from the root عمل [3ml].

In fact, Arabic words are agglutinated words, composed by an inflected word form (base) and attachable clitics. Four categories of clitics were presented in [23]. As the Arabic sentence is written from right to left, the segmentation of an Arabic word can be composed like this (Table 2).

A word in Arabic, can correspond to multiple English words, where the English words are spread at distinct places in the English sentence. This problem complicates the alignment between an Arabic and an English sentences and increases the OOV

Table 1 Example of Arabic words derived from the root “3ml” [13]

Pattern	Arabic word	English Translation
عَمل	فَعل [Eaml]	job
عامِل	فاعِل [EAmil]	worker
عَمَلَ	مَفعل[Eamala]	worked
مَعمل	فَعل [maEml]	workshop
عُمِلَ	فُعِلَ [Eumila]	has been worked

Table 2 Segments of an Arabic word

Enclitic	Suffix	**BASE**	Prefix	Proclitic

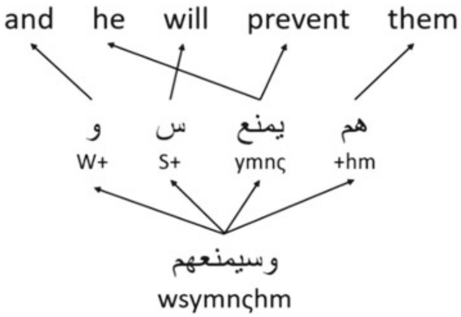

Fig. 1 Illustration of word alignment between a sentence in English and its translation in Arabic [43]

(out-of-vocabulary) rate in a SMT system [39]. The example presented in Fig. 1 illustrates the alignment problem between an agglutinated Arabic word and an English sentence.

As illustrated by the Fig. 1, a single Arabic word can correspond to multiple English words. Hence, to decrease token sparsity and improve the alignment in SMT systems, we lead to the morphological segmentation step of the Arabic text. The segmentation amounts to separating the inflected base from the clitics attached to it. This pre-processing step proves it efficiency in many NLP applications dealing with Arabic language [39]. The segmentation of Arabic words into correct morphemes is a challenging task. The segmentation improves the evaluation scores of the SMT system [39].

2.2 *Arabic Language in Microblogs*

As presented in the previous section, the complexity of the morphology together with the underspecification of the orthography in Arabic language create a high degree of ambiguity. This ambiguity increases more and more in the social media texts. Nowadays, there are many other issues that appeared with the spread of social media platforms. The texts in microblogs, for example Twitter, are short, noisy and written in a non-standard orthography style.

Indeed, social media users tend to commit spelling defects; They tend to write words in Arabic using the Latin alphabets and numbers. Also, each user transliterates the word in its own way. This phenomenon is called “Arabizi” [1, 3, 7]. For example, the Arabic letter ‘ح’ [H] is often transliterated by the number ‘7’, the letter ‘ق’ [q] by the number ‘9’, etc. In tweets, the users transliterate the proper name “احمد” [AHmd] in different ways: *ahmed*, *ahmad*, *ahmd*, *a7mad*, *a7med*, *a7mmd*, *a7md* or *ahmmd*.

Likewise, the proper name "أشرف" [a$rf], is written as *ashraf*, *ashref*, *ashrf*, *shrf*, *achraf*, or *aschraf* [32].

Moreover, users on social media tend to use more than one language in the text they publish. They often alternate between their dialect and a foreign language. This phenomenon is called *code-switching* and it has been studied in [2]. This problem greatly affects applications that handle only one language.

The use of dialectal Arabic (DA) in social media is a serious problem for Arabic NLP applications. Firstly, the current NLP tools for DA are not able to handle the dialects used in social media texts because they lack strict writing standards. For example, the MSA phrase "لا يلعب" [lA ylEb] can be written in Egyptian dialect in different ways like "مابيلعبش" [mAbylEb$], "مايلعبش" [mAylEb$], "ميلعبش" [mylEb$], "مابلعبش" [mAblE$]. Also they can be transliterated in different ways by many users like *Mayel3absh, mabyelaabsh, mabyel3absh, etc.* [7].

Messages published on Twitter (tweets) are short, noisy and they have a rich structure. It contains different special fields like the *username*, *hashtags*, *retweet*, etc. In fact, *usernames* were the subject of a named entities processing applications for Arabic in [32]. Also *hashtags* were the purpose of an improved machine translation application in [14].

3 Overview of Statistical Machine Translation

Statistical Translation model has been proven in machine translation (MT) applications since 1990. Inspired by the noisy channel model of Claude Shannon [44], Brown and his team at IBM proposed the first Statistical Translation System (SMT) [4, 5]. The authors supposed that any sentence in a source language S, can be translated into another sentence in target language T. Thus for any pair of sentences (s, t), they assign a probability of translation $p(t|s)$. This probability is interpreted as a feature that the machine translation will perform the translation hypothesis $\hat{t}$ in the target language T given a sentence s in the source language S. The problem of SMT aims to maximize the probabilities between the translation model and the language model and choose the best translation hypothesis $\hat{t}$. Hence, applying Bayes' theorem, mathematically, the problem is described as below:

$$p(t|s) = \frac{p(s|t)\, p(t)}{p(s)} \tag{1}$$

Because the probability $p(s)$ is independent to the source sentence t, so the equation (1) is simplified as follow:

$$\hat{t} = argmax_{t \in t^*}\, p(s|t)\, p(t) \tag{2}$$

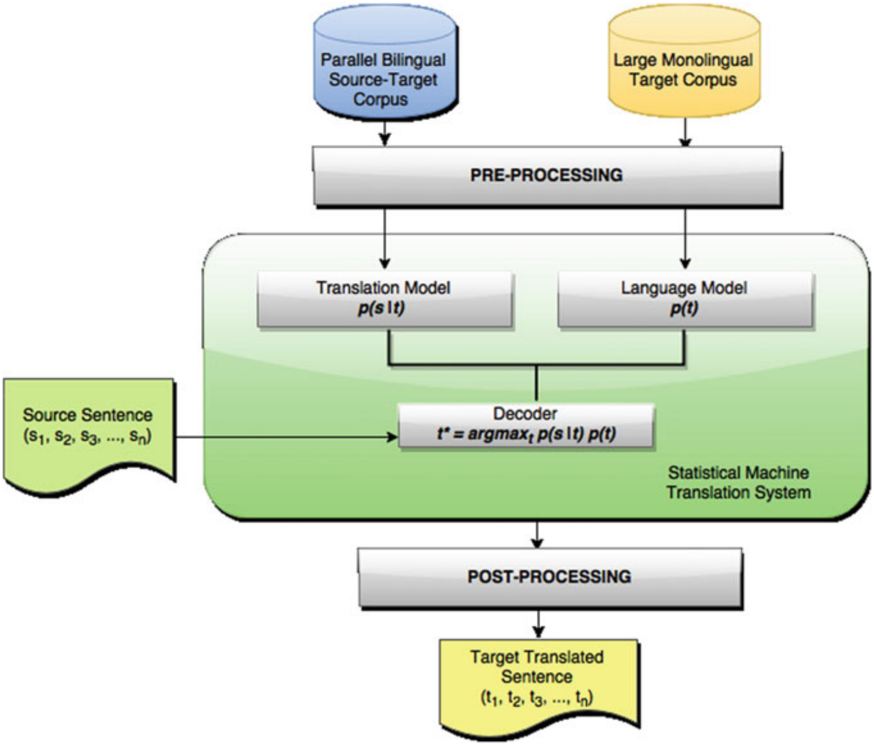
Parallel Bilingual Source-Target Corpus
Large Monolingual Target Corpus
PRE-PROCESSING
Translation Model $p(s|t)$
Language Model $p(t)$
Source Sentence $(s_1, s_2, s_3, ..., s_n)$
Decoder $t^* = argmax_t\, p(s|t)\, p(t)$
Statistical Machine Translation System
POST-PROCESSING
Target Translated Sentence $(t_1, t_2, t_3, ..., t_n)$

Fig. 2 General architecture of the state of the art of statistical machine translation based system

As described in the Eq. (2), the architecture of a SMT-based system is composed of two important components: the **translation model (TM)** $p(s|t)$ and the **language model (LM)** $p(t)$. In fact, the translation model (TM) contains the list of phrases translations, after the training phase by using a large parallel bilingual corpus and the language model which is built from a monolingual corpus in the target language. The translation model gives the best translation hypothesis according to the source input text while the language model ensures that this hypothesis is syntactically correct for the target text regardless of the source input text.

Moreover, in the architecture of a SMT-based system, there is also a third important component, the **decoder**. It aims to search and to find out the best translation hypothesis $\hat{t}$ among all possibilities proposed by the system. In a non exhaustive list, there are many decoders in the literature such as Pharaoh[3] [25], Portage [22] and Moses[4] [26]. The Fig. 2 describes the general architecture of the state of the art of SMT-based system.

[3] http://www.isi.edu/licensed-sw/pharaoh/.

[4] http://www.statmt.org/moses.

Actually, many SMT-based systems perform translation models either based on words or based on phrases. The first approach is known as the Word-Based Statistical Machine Translation (WBSMT) in which the system is based on an automatic word-to-word alignment [34]. The second approach is known as the Phrases-Based Statistical Machine Translation (PBSMT), in which the system considers the alignment unit as a contiguous sequences of words or segment [27].

3.1 Language Model

The language models (LM) are widely used in NLP for several applications, especially machine translation and speech recognition. Thus, the language model is considered as a basic component in SMT systems. It aims to estimate the probabilities that a phrase or a sequence of words appear in the target language [4]. The language models in the SMT systems are built based on a monolingual corpus in the target language. They allow to calculate the likelihood of the translation hypotheses in the target language. Mathematically, the language model $p(t)$ for the sentence t, which is composed by n words $t = w_1\ w_2\ w_3\ \dots\ w_n$, is defined as follows:

$$p(t) = \prod_{i=1}^{n} p(w_i|w_1 \dots\ w_{i-1}) \tag{3}$$

To simplify this modeling, we assume that the w_i words of t depend only on the previous $(i-1)$ words. We can therefore reformulate the Eq. (3) as follows:

$$p(t) = p(w_1)\,p(w_2|w_1)\,p(w_3|w_1w_2)\ \dots\ p(w_1|w_1w_2\ \dots\ w_{n-2}w_{n-1}) \tag{4}$$

This language model is known as the n-gram model. It makes predictions based on a fixed size search window, containing n words. Hence, for each word, a probability is calculated by taking into account the $(n-1)$ words that precede the current word in the target sentence. This probability represents the dependency of each word with respect to the $(n-1)$ words that precede it, as indicated by the Eq. (4). Statistically, if the sequence of words to be translated does not exist in the language model, its probability will be null.

3.2 Word Alignment

Automatic word alignment remains an important component for all SMT approaches. Given a bilingual sentence pair, the general definition of word alignment refers to any defines set of links between lexical units that are translations of each other. In 1990, the first probabilistic models for machine translation was based on words, i.e. the translation unit that appears in the probability laws is the word [4]. An example of word alignment of a sentence in English and its translation in Arabic was illustrated in Fig. 1.

The first challenge in the word-based SMT models consists of establishing the mapping between words in the source sentence and words in the target sentence. In this modeling problem, a hidden variable a is used to account all the possible pair-wise alignment links between both sentences. This alignment problem is mathematically described, between the source sentence and the target sentence, as below:

$$p(s|t) = \sum_{a} p(s, a|t) \tag{5}$$

$p(s, a|t)$ is generally expressed as:

$$p(s, a|t) = p(J|t) \sum_{j=1}^{J} p(a_j|s_1^{j-1},\ a_1^{j-1}, J, t)\, p(s_j|a_1^{j-1},\ s_1^{j-1}, J, t) \tag{6}$$

where,

J: length of the source sentence s
s_j: word in position j of source sentence s
a_j: hidden alignment of word s_j indicating the position at which s_j aligns in the target sentence t.

The Eq. (6) aims to generate an alignment between a source sentence and a target sentence. Firstly, the length J of the source sentence s is manually chosen, given what the target sentence is known. The choice of the position to link the first source word, given the target sentence and the length of the source sentence, can be made. Then the identity of the first source word is chosen, given the target sentence, the length of the source sentence and the target word linked to the first source position, and so on.

Five models, known as "*IBM models*" from 1 to 5, have been proposed by [4, 5]. They aim to maximize the translation probability $P(t|s)$ of a target sentence given a source sentence.

- *IBM 1*: This model is a simple lexical translation model, which makes use of co-occurrence of word pairs. It assigns only lexicon probabilities.
- *IBM 2*: This model adds absolute reordering probabilities by introducing local dependencies.
- *IBM 3*: This model adds fertility model, which depends only on the source sentence.
- *IBM 4*: This model takes into account the relative reordering probabilities.
- *IBM 5*: This model fixes deficiency. It limits the waste of probabilities mass on impossible situation. It is considered as the non-deficient version of *IBM 4*.

These models did not have much success due to their several disadvantages. They allow only the one-to-many words alignment and the alignment mapping is restricted to source-to-target locations. The IBM models assume that the words are syntactically independent of one another. Hence, the context of the text to be translated is not taken into account, and the models can generate a confusion in the meaning of the sentence to be translated, for example, in case of polysemous words. For example, the word "*livre*" in French, could be translated in English such as a "*book*" or a "*pound*" [10]. For these reasons, most current SMT systems are not based on word-to-word approaches but based on phras-based approaches.

3.3 Translation Model

Most phrases-based translation models are commonly used by the research community in the machine translation domain [27]. A phrase can be a word or a set of words. Actually, a sentence is tokenized and then segmented into many phrases. The one-to-one and one-to-many words alignments are offered by the IBM models while many-to-many words alignments are offered by phrases-based models. Once the alignments are established, a probability score is calculated for all phrases. Each source phrase can have several translation hypotheses in the target language. Then the choice of the candidate phrase is based on the probabilities stored in the *phrases translation table*.

The translation probability $\phi(\bar{s}|\bar{t})$ is defined with score by relative frequency as follows:

$$\phi(\bar{s}|\bar{t}) = \frac{count(\bar{s},\bar{t})}{\sum_{\bar{t}_i} count(\bar{s},\bar{t}_i)} \tag{7}$$

3.4 Decoding

The translation process of a source sentence into a target sentence is known as decoding. This process consists of three following steps: (1) to segment the source sentence into phrases, (2) to translate the source phrases according to the probabilities in the translation model, (3) to reorder the source phrases according to the target language. The decoding process in phrases-based SMT is defined as follows [27]:

$$p(\bar{s}_i|\bar{t}_i) = \prod_{i=1}^{n} \phi(\bar{s}_i|\bar{t}_i)\ d(start_i - end_{i-1} - 1) \tag{8}$$

where:

- $\bar{s}_i$: a set of phrases of the source sentence s
- $\bar{t}_i$: translation of each source phrase $\bar{s}_i$
- $\phi(\bar{s}_i|\bar{t}_i)$: phrase probability according to the translation table
- $d(start_i - end_{i-1} - 1)$: reordering probability of each target phrase $\bar{t}_i$.

The decoding problem is known as a NP-complete problem [24]. To solve this problem, it is very important to reduce the search space, when effective solutions are searched. Wang and Waibel [47] have proposed a stack-based decoder based on the A* search algorithm to find out the best hypothesis. Other researchers used weighted finite states transducers to implement an alignment model [49]. In [12] the authors transformed the decoding problem into a linear programming problem by implementing a beam search algorithm. Also, dynamic programming algorithms with pruning have been implemented in [25, 37]. Koehn et al. [26] have presented the open source Moses decoder which becomes the state-of-the-art used in the SMT research community.

4 Translating Arabic Tweets

The SMT methods depend on the quantity and the quality of the data used for the translation model (TM) and the language model (LM). It's very important that the training set and the test set remain in the same domain (for example the domain of news, medicine, social media, etc.). This gives a more efficient and robust MT system [28].

Although, this is not evident in our work because the parallel corpus for Arabic-English tweets is not available. For this reason, our method is to adapt a state of the art SMT system the most possible, to be able to translate tweets in an efficient way. So, the training data set is an out-of-domain parallel corpus in Modern Standard Arabic (MSA). Inspired by the work in [21], the first strategy is to adapt the system by incorporating a big LM for tweets (an in-domain language model). The second strategy is tuning the SMT with an in-domain data.

4.1 Data Collection

Although, this is not evident in our work as there is no parallel corpus available for Arabic-English tweets. Therefore, our method focuses on the adaptation of a state-of-the-art SMT system to be able to translate tweets in an efficient way. So, the training data set is an out-of-domain parallel corpus in MSA. The first strategy aims to adapt the system by incorporating a big LM for tweets (an in-domain language model). The second strategy uses SMT tuning with an in-domain data [21].

4.2 Data Collection

For the first strategy, we need a large corpus of tweets in English to train the LM, and a small corpus in Arabic for the test set. To do so, we collected tweets via

Table 3 Collected tweets

Language	Number of tweets
Anglais	255 602
Arabe	1 930

the Streaming API of Twitter,[5] using Twitter4j which is an open source library for java [6] [48].

Influenced by the political events in the Arab world, we chose the following list of Arabic keywords to collect tweets in Arabic for the test set: (“الجزيرة” [Aljzyrp], “سوريا” [swryA], “حلب” [Hlb], “بشار” [b$Ar], “ثورة” [vwrp], “الربيع العربي” [AlrbyE AlErby]). Also we adjusted the library to download only tweets written in Arabic.

After that, we used the translations of Arabic keywords to collect English tweets, to generate the LM: (*Syria, Halab, Bachar, revolution, arab spring, AlJazeera*). We collected 255,602 English tweets and 551 Arabic tweets, during four months, from September 2015 to December 2015. The collected tweets are described in the Table 3):

The collected data were very noisy and contained many orthographic and grammar mistakes. Moreover, the tweets follow a standard form and use special terms distinguishing them from other microblogs:

- The *Username* which identified each user and it is preceded by the symbol @ (@UserName).
- The *hashtag* is preceded by the character (#). It highlights the keywords to a specific event. By clicking on a given *hashtag*, tweets that contain the same *hashtag* appear.
- The *retweet* start by (RT; @username). It indicates that the same tweet is republished by another user.
- The URL which is inserted in the end of a tweet.

In this paper, we are interested only in the raw text of the tweet, or the message itself. Thus, we deleted all the specific fields of Twitter listed previously.

Pre-processing English Tweets When we analyzed the collected tweets, we observed that users tend to use abbreviations instead of a whole sentence. For example, the sentence “*just for you*” is written as “*just4u*”. Also to express their feelings, users use the stretched words like “*haaaaappy*” instead of “*happy*”.

To deal with this problem, we used a dictionary of non-standard words and their standards correspondents. This dictionary was proposed at a shared task organized

[5] https://dev.twitter.com/streaming/public.

[6] http://twitter4j.org/en/index.html.

as part of the W-NUT-2015 (ACL 2015 Workshop on Noisy User-generated Text [W-NUT])[7] [19, 36].

Once the lexical normalization step was done, we obtained a corpus of tweets in English ready for building the LM using the SRILM toolkit [45].

Pre-processing Arabic Tweets This corpus was considered as the reference to evaluate the translation systems. After the pre-processing and the normalization steps, the corpus was translated into English by a human expert.

Evidently, many lexical and orthographic errors were found and were processed. For examples, compound expressions formed by two or more words separated by an underscore such as “الربيع_العربي” [AlrbyE_AlErby], or “تنظيم_الدولة” [tnZym_Al dwlp]. The Stretched words (examples: “تكلنـــــــي” [tklmn______y], “دبـــــــــي” [db___________y]) and also words with repeated letters (examples: “كبرررررر” [kbrrrrrr], “لايمككننا” [lAymkknnA] had a significant high presence in the collected data set.

The dialectical Arabic, especially the levantine dialect, were frequently used. However, we did not deal with this type of words. Only the lexical and orthographic errors presented before/above have been dealt with in the normalization step.

We applied the state-of-the-art pre-processing steps for Arabic, such the orthographic normalization for the letters “Hamza” [a] and “Ya” [y]. These two letters are inconsistently spelled using different forms. The “Hamza” has different forms like (‘أ’ , ‘آ’, ‘إ’). In this paper we represent it in a single way as bare Alif ‘ا’ [A]. Also, the letter “Ya” [y] is written with dots or dotless. All the letters ‘ي’ [y] were normalized by the letter ‘ى’ [Y], the dotless form of “Ya” [y], to reduce the degree of the spelling ambiguity.

In the next step, we applied a morphological segmentation of Arabic words using the morphological analyzer MADA [15]. This task was deeply studied in many NLP applications for Arabic and it generally improves the state-of-the-art baseline systems in terms of BLEU score [16]. Many segmentation scheme were proposed in literature for Arabic: S1, ON, D1, D2, D3, WA, TB, MR, L1, L2 et EN [16, 39]. In this study, we used the D2 segmentation scheme, which its efficiency for MT applications is well known. In the D2 configuration considered here, four proclitic particles (l+ (to/f or), b+ (by/with), k+ (as/such), and f +(in)) and one conjunction proclitic (w+ (and)) are identified. These are separated from their associated root words.

The normalization of the letters “Hamza” [a] and “Ya” [y] and the segmentation were also applied for Arabic in the parallel corpus.

[7] The dictionary contains 44983 pairs of words, available on-line at: http://noisy-text.github.io/2015/norm-shared-task.html.

Table 4 Details of the parallel corpus

	Size (Mo)	Words	Tokens
Arabic (source)	165,5	16 866 817	18 791 118
English (target)	113,5	18 608 307	19 408 007

4.3 Experiments

Data To build the SMT for Arabic-English tweets, we use the parallel corpus from the United Nation (UN).[8] The training corpus was about 280 mega-octet (Mo). The Arabic side of the parallel corpus contains about 16 million words and 18 million tokens after the segmentation and the normalization steps. The Table 4 presents the detailed statistics for both languages: Arabic and English.

In the experiments, we use three LMs: one LM of tweets (presented in the Sect. 1.4, which is crawled from Twitter), a big LM of tweets used in [6], and a LM in MSA (the target side of the parallel corpus). The LMs are 3-gram LM, generated with the SRILM toolkit [45]. The pre-processing steps, i.e. the normalization of the letters “Hamza” [a], “Ya” [y] and the segmentation, were applied to the Arabic data (LM, parallel data, and test).

For the tuning step, we used a small parallel corpus with around 1000 sentences extracted from the UN corpus. Secondly, we tuned the translation systems using a small in-domain parallel corpus (Arabic-English tweets used in [38]).

Experimental Setup The experiments were carried out using the open source phrase-based SMT system Moses [26], with maximum phrase length of 10.[9] To obtain the word-to-word-alignments during the training between Arabic and English, we used MGIZA [11], a multi-threading version of GIZA++ [34].

Finally, the feature weights were estimated using Moses built-in minimum error rate training procedure (MERT) [33], which uses two different types of tuning sets: MSA and tweets.

Many systems were trained according to defined pre-processing models. We evaluated those systems in terms of BLEU [35], and the out of vocabulary words rate (OOV).

4.4 Results

Table 5 shows different results of the SMT systems for tweets, which was trained using an out-of-domain parallel corpus. To better adapt our system to the domain of social media, especially the tweets, we try to test all the possible combinations

[8] The UN parallel corpus is available at: http://www.un.org/en/documents/ods/.

[9] http://www.statmt.org/moses/?n=FactoredTraining.PrepareTraining.

Table 5 Overview of experiments and results for Arabic-English tweet's MT

Systems	LM	Pre-processing	Tuning	Results	
				BLEU	OOV
1	tweets	–	UN	2.39	28.74
2	tweets	+	UN	6.09	9.95
3	tweets	+	tweets	10.31	7.16
4	UN	–	UN	6.31	24.53
5	UN	+	UN	8.39	12.11
6	UN	+	tweets	8.96	9.61
7	tweets+UN	–	UN	2.49	27.75
8	tweets+UN	+	UN	3.50	11.18
9	tweets+UN	+	tweets	**10.98**	6.89
10	*BIG_tweets*	+	tweets	10.58	7.49

(LM and tuning set). Also, we tested the baseline systems without pre-processing the parallel and the test data. Systems 1, 4 and 7 are considered as baseline systems without any pre-processing strategy (−). All the other experiments (system 2, 3, 5, 6, 8, 9 and 10) were carried out after completing the pre-processing steps (+).

The best system has reached a BLEU score of 10.98. This result remains weaker than the result of a system with training test and dev corpus belong to the same domain, as in [39]. The pre-processing is very important and it has improved the BLEU scores by 1 to 4 points and reduced the OOV rate. As a final step, we tuned the systems using an in-domain corpus, which improved the BLEU scores and the OOV rates. The best combination resulting from these experiments is therefore a system based on two LMs (UN and tweets) with the necessary pre-processing steps for the training, tuning and test corpus.

4.5 Discussion

The presented results prove that for the SMT methods, the domain of the training and the test data is very important. Thus, when we combined the LM of MSA and the LM of tweets with an in-domain tuning set, we obtained the best result in term of BLEU score.

However, translation systems were not able to translate many named entities found in the test set like "بوتين" [bwtyn], "درعا" [drEA], "الجزيرة" [Aljzyrp], "تويتر" [twytr], "جوجل" [jwjl], "حماه" [HmAh], "حمص" [HmS], "داعش" [dAE$], "روسيا" [rwsyA].

Also, we noted that DA such as the levantine was used extensively in the collected tweets. As the parallel corpus is in MSA, many levantine expressions were not matched.

For the various works related to MT for microblogs, the BLEU scores are often low and despite the pre-processing performed, the quality of translations of this type of data remains poor. In [30] the BLEU score was 8.75, with out-of-domain parallel corpus. In the same context, [46] obtained 22.57 of BLEU score translating tweets from Spanish to Basque. This result were expected as the two languages are semantically very closed. Also, In [20], the score of the tweet's MT from German to English, was 15.68.

In an advanced comparison study, we tested our translation model with an in domain-data (MSA test set), We used the NIST evaluation set, MT05. The resulted BLEU score was 16.91 and the OOV rate was 3.51. The BLEU outperforms the translation of tweets for the following reasons: the test set is not noisy like tweets, and the parallel corpus and the test corpus are in the same domain.

5 Conclusion and Future Work

In our work, we have explored some of the problems facing (Arabic-English) tweets translation. Thus, we mounted a statistical machine translation using the most popular decoder for segment-based statistical machine translation systems, Moses [26].

Statistical machine translation (SMT) systems are based on a parallel corpus and on a monolingual corpus to train a language model. In our work, the parallel corpus of tweets for the Arabic-English language pair was not available. To overcome this problem, we used several adaptation strategies for the translation systems for tweets. These strategies are based on the use of a big language model for tweets, and a tuning step with a dev in-domain corpus.

Dealing with Arabic language, a morphological rich and complex language, we perform a pre-processing step including the normalisation of the letters "Hamza [a]" and "Ya [y]" and the segmentation of Arabic words. This step was very important to reduce the degree of sparsity and improve the alignment between Arabic and English words.

Also, we carried out the spelling and orthographic mistakes in tweets, by normalising the stretched words, the transliterated expression, etc. These pre-processing steps were very helpful and ameliorate the BLEU score, which reach 10.98.

In the future work, it will be useful to enlarge the parallel corpus by a parallel data, from Twitter or microblogs in general. This will improve the results. Also, dealing with the problem of "arabizi" and "code switching" in tweets is very important, and could improve the quality of the translation system for Arabic tweets.

References

1. Adouane, W., Semmar, N., Johansson, R., Bobicev, V.: Automatic detection of Arabicized Berber and Arabic varieties. VarDial **3**, 63 (2016)
2. Barman, U., Das, A., Wagner, J., Foster, J.: Code mixing: a challenge for language identification in the language of social media. In: First Workshop on Computational Approaches to Code Switching (EMNLP 2014), pp. 13–23. Association for Computational Linguistics (ACL) (2014)
3. Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., Rambow, O.: Transliteration of Arabizi into Arabic orthography: developing a parallel annotated Arabizi-Arabic script SMS/chat corpus. In: Workshop on Arabic Natural Langauge Processing (ANLP), pp. 93–103 (2014)
4. Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Comput. Linguist. **16**(2), 79–85 (1990)
5. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
6. Cherry, C., Guo, H.: The unreasonable effectiveness of word representations for twitter named entity recognition. In: HLT-NAACL, pp. 735–745 (2015)
7. Darwish, K.: Arabizi Detection and Conversion to Arabic, pp. 217–224. Association for Computational Linguistics, Doha, Qatar (2014)
8. Dridi, H.E.: Détection d'évènements à partir de Twitter. Ph.D. thesis, Université de Montréal (2015)
9. Farzindar, A., Roche, M.: Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues. Traitement Automatique des Langues **54**(3), 7–16 (2013)
10. Gahbiche-Braham, S.: Amélioration des systèmes de traduction par analyse linguistique et thématique. Ph.D. thesis, Université Paris Sud (2013)
11. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pp. 49–57. Association for Computational Linguistics (ACL) (2008)
12. Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K.: Fast decoding and optimal decoding for machine translation. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 228–235. Association for Computational Linguistics (ACL) (2001)
13. Ghoul, D.: Outils génériques pour l'étiquetage morphosyntaxique de la langue arabe: segmentation et corpus d'entraînement (2011)
14. Gotti, F., Langlais, P., Farzindar, A.: Translating government agencies tweet feeds: specificities, problems and (a few) solutions. In: The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013), p. 80 (2013)
15. Habash, N., Rambow, O.: Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 573–580. Association for Computational Linguistics (2005)
16. Habash, N., Sadat, F.: Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 49–52. Association for Computational Linguistics (ACL) (2006)
17. Habash, N., Sadat, F.: Challenges for Arabic machine translation. In: Abdelhadi Soudi, Ali Farghaly, Günter Neumann, Rabih Zbib (eds.) Natural Language Processing, pp. 73–94. Amsterdam (2012)
18. Habash, N.Y.: Introduction to Arabic natural language processing. Synth. Lect. Hum. Lang. Technol. **3**(1), 1–187 (2010)
19. Han, B., Cook, P., Baldwin, T.: Lexical normalization for social media text. Assoc. Comput. Mach. Trans. Intell. Syst. Technol. (TIST) **4**(1), 5 (2013)

20. Jehl, L., Hieber, F., Riezler, S.: Twitter translation using translation-based cross-lingual retrieval. In: Proceedings of the Seventh Workshop on Statistical Machine Translation, pp. 410–421. Association for Computational Linguistics (ACL) (2012)
21. Jehl, L.E.: Machine Translation for Twitter. Master's thesis, Speech and Language Processing School of Philosophy, Psychology and Language Studies, University of Edinburgh (2010)
22. Johnson, J.H., Sadat, F., Foster, G., Kuhn, R., Simard, M., Joanis, E., Larkin, S.: Portage: with smoothed phrase tables and segment choice models. In: The Workshop on Statistical Machine Translation, pp. 134–137. Association for Computational Linguistics, New York City (2006)
23. Kadri, Y., Nie, J.Y.: Effective stemming for Arabic information retrieval. In: The Challenge of Arabic for Natural Language Processing/Machine Translation NLP/MT, pp. 68–74 (2006)
24. Knight, K., Marcu, D.: Machine translation in the year 2004. In: ICASSP (ed.) ICASSP (5) International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 965–968. Institute of Electrical and Electronics Engineers (IEEE) (2005)
25. Koehn, P.: Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In: Conference of the Association for Machine Translation in the Americas, pp. 115–124. Springer (2004)
26. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics (ACL) (2007)
27. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 48–54. Association for Computational Linguistics (ACL) (2003)
28. Langlais, P., Gotti, F., Patry, A.: De la chambre des communes à la chambre d'isolement: adaptabilité d'un système de traduction basée sur les segments. In: Les actes de TALN, pp. 217–226 (2006)
29. Le, N.T., Mallek, F., Sadat, F.: UQAM-NTL: named entity recognition in twitter messages. WNUT **2016**, 197 (2016)
30. Ling, W., Xiang, G., Dyer, C., Black, A.W., Trancoso, I.: Microblogs as parallel corpora. Assoc. Comput. Linguist. (ACL) **1**, 176–186 (2013)
31. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How translation alters sentiment. J. Artif. Intell. Res. (JAIR) **55**, 95–130 (2016)
32. Mubarak, H., Abdelali, A.: Arabic to English person name transliteration using Twitter. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Slovenia (2016)
33. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 160–167. Association for Computational Linguistics (2003)
34. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**(1), 19–51 (2003)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for computational linguistics, pp. 311–318. Association for Computational Linguistics (2002)
36. Pennell, D.L., Liu, Y.: Normalization of informal text. Comput. Speech Lang. **28**(1), 256–277 (2014)
37. Quirk, C., Moore, R.: Faster beam-search decoding for phrasal statistical machine translation. Machine Translation Summit XI (2007)
38. Refaee, E., Rieser, V.: Benchmarking machine translated sentiment analysis for Arabic tweets. In: Student Research Workshop (SRW-2015), pp. 71–78 (2015)
39. Sadat, F., Habash, N.: Combination of Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1–8. Association for Computational Linguistics, Sydney, July 2006

40. Sadat, F., Kazemi, F., Farzindar, A.: Automatic identification of Arabic language varieties and dialects in social media. In: The 4th International Workshop on Natural Language Processing for Social Media of (SocialNLP 2014) (2014)
41. Sadat, F., Mallek, F., Sellami, R., Boudabous, M.M., Farzindar, A.: Collaboratively constructed linguistic resources for language variants and their exploitation in NLP applications—the case of Tunisian Arabic and the social media. In: Workshop on Lexical and Grammatical Resources for Language Processing, p. 102. Citeseer (2014)
42. Salameh, M., Mohammad, S.M., Kiritchenko, S.: Sentiment after translation: a case-study on Arabic social media posts. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL), pp. 767–777. Association for Computational Linguistics, May 2015
43. Salameh, M.K.: Morphological solutions for Arabic statistical machine translation and sentiment analysis. Ph.D. thesis, University of Alberta (2016)
44. Shannon, C.E.: The Mathematical Theory of Communication. Urbana (1949)
45. Stolcke, A., et al.: Srilm—an extensible language modeling toolkit. In: ICSLP 2, pp. 901–904, Sept 2002
46. Toral, A., Wu, X., Pirinen, T., Qiu, Z., Bicici, E., Du, J.: Dublin City University at the TweetMT 2015 shared task. In: Tweet Translation Workshop at the International Conference of the Spanish Society For Natural Language (SEPLN 2015) (2015)
47. Wang, Y.Y., Waibel, A.: Decoding algorithm in statistical machine translation. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics, pp. 366–372. Association for Computational Linguistics (1997)
48. Yamamoto, Y.: Twitter4J—an open-sourced, mavenized and Google App Engine safe Java library for the Twitter API, released under the BSD license (2009)
49. Zhang, M., Li, H., Kumaran, A., Liu, M.: Report of news 2012 machine transliteration shared task. In: Proceedings of the 4th Named Entity Workshop, pp. 10–20. Association for Computational Linguistics (2012)