Machine translation for Arabic dialects (survey)

Salima Harrat, Karima Meftouh, Kamel Smaïli

Machine Translation for Arabic Dialects (Survey)

Salima Harrat[a], Karima Meftouh[b], Kamel Smaili[c]

[a]*École Supérieure d'Informatique (ESI), École Normale Supérieure de Bouzaréah (ENSB), Algeria*
[b]*Badji Mokhtar University, Annaba, Algeria*
[c]*Campus Scientifique LORIA, France*

Abstract

Arabic dialects also called colloquial Arabic or vernaculars are spoken varieties of Standard Arabic. These dialects have mixed form with many variations due to the influence of ancient local tongues and other languages like European ones. Many of these dialects are mutually incomprehensible. Arabic dialects were not written until recently and were used only in a speech form. Nowadays, with the advent of the internet and mobile telephony technologies, these dialects are increasingly used in a written form. Indeed, this kind of communication brought everyday conversations to a written format. This allows Arab people to use their dialects, which are their actual native languages for expressing their opinion on social media, for chatting, texting, etc. This growing use opens new research direction for Arabic natural language processing (NLP). We focus, in this paper, on machine translation in the context of Arabic dialects. We provide a survey of recent research in this area. We report for each study a detailed description of the adopted approach and we give its most relevant contribution.

Keywords: Arabic dialect, Modern Standard Arabic, Machine translation

1. Introduction

Arabic dialects are informal spoken language used all over Arab countries. These dialects are used in everyday life, in contrast to modern standard Arabic that is used in official speeches, newspapers, school, etc. This coexistence of two variants of a language in the same community is known as diglossia which is defined in (Ferguson, 1959) as: "A relatively stable language situation in which, in addition to the primary dialects of the language, there is a very divergent highly codified superposed variety, the vehicle of a large and respected body of written literature which is learned largely by formal education and is used for most written and formal spoken purposes, but is not used by any sector of the community for ordinary conversation". This linguistic phenomenon exists in all Arab countries. Furthermore, in the last decade these dialects emerged in social networks, SMS, TV-Shows, etc. They are increasingly used even in a written form. This usage generates new needs in NLP area. Indeed, these dialects are not enough resourced in terms of NLP tools and those concerning modern standard Arabic (MSA) are not adapted to process them.

In this paper, we focus on machine translation of Arabic dialects. This area has become an interesting research field because of the many challenges to overcome. In fact, Arabic dialects differ from Standard Arabic at phonological, lexical, morphological and syntactic levels. They simplify a wide range of written Arabic rules[1] on one hand but add other new rules[2] that generates a lot of complexities on the other hand. In addition, these dialects (especially Maghrebi ones) are influenced by other languages such as French, Spanish, Turkish and Berber. Besides the fact that these dialects are different from Standard Arabic, they are also different to each other; even within the same country these dialects are not the same.

2. NLP challenges for Arabic dialects

Arabic dialects, despite their large use are under-resourced languages, they lack basic NLP tools. Except some work dedicated to Middle-east dialects (Egyptian dialect mostly), these dialects are not enough studied regarding to NLP area. Most MSA resources and tools are not adapted to them and do not take into account their features. The reader can refer to (Habash, 2010) which presented a comprehensive survey on Arabic NLP, the work focused on MSA but included many interesting notes on pratical issues

[1] The dual form, for example, as well as the feminine plural form used in standard Arabic do not exist in most Arabic dialects.

[2] Standard Arabic has a strong case system where most cases are denoted by diacritics. In Arabic dialects, there is no grammatical case, thus generating more syntactic ambiguities compared to MSA. Also, the verbs negation in dialects is more complex than MSA, the circumfix negation is placed around the verb with all its prefixes and suffixed direct and indirect object pronouns.

concerning Arabic dialects or to (Shoufan and Al-Ameri, 2015) where authors reported all available tools and resources recently produced for these dialects. One of the main issue of Arabic dialects is the fact that they have no conventional orthographies for writing them. Their large use (because the advent of Internet technologies) produces important volumes of data which are difficult to exploit and require important pre-processing steps.

In the area of machine translation, Arabic dialect translation research efforts are still at an early stage. Rule-based approaches are difficult to envisage because of unavailability of dedicated tools for most of these dialects. Indeed, these approaches are being used less and less in MT systems because they are time consuming and require important linguistic resources. Also, MT systems based on these approaches are difficult to maintain, adding new linguistic features involves updating rules or adding new rules. For Arabic dialects, these approaches are more problematic. These dialects are not written and have no strong theoretical linguistic studies that could allow such approach. In addition, these dialects differ from one Arab country to another, even in the same country significant variations exist, any rule-based MT system could not take into account all related features. On the other hand, data-driven approaches are also hard to consider due to the lack of resources like parallel and even monolingual corpora. In the context of statistical machine translation, Arabic dialects lack bi-texts with reasonable sizes that allow building efficient statistical machine translation (SMT) systems readily.

It should be noted that this issue does not arise only in the case of Arabic dialects; it concerns also several other under-resourced languages and many research activities focus on machine translation in the context of under-resourced or non-resourced languages. The main idea of these contributions is exploiting the proximity between an under-resourced language and the closest related resourced language (Cantonese⇒Mandarin (Zhang, 1998), Czech⇒Slovak (Hajič et al., 2000), Turkish⇒Crimean Tatar (Altintas and Cicekli, 2002), Irish⇒Scottish Gaelic (Scannell, 2006), Indonesian⇒English using Malay (Nakov and Ng, 2012) and Standard Austrian German⇒Viennese dialect (Haddow et al., 2013)).

3. Machine translation related to Arabic dialects

In this section we present most important studies dedicated to Arabic dialects machine translation. We first introduce research dedicated to machine translation between modern standard Arabic and its dialects. Then, we focus on MT between foreign languages and Arabic dialects. In this context, we would point out that all contributions concern mainly English language (as we will see later). We attempt to draw a clear picture of each study by describing its approach, the used data and the achieved results. We will show that most of them exploit the proximity between these dialects and MSA, and attempt to use available MSA resources to deal with Arabic dialects.

3.1. Translating between MSA and Arabic dialects

Bakr et al. (2008) presented a generic approach for converting an Egyptian colloquial Arabic sentence into vocalized MSA sentence. They combined a statistical approach to automatically tokenize and tag Arabic sentences and a rule-based approach for creating the target diacritized MSA sentence. The work was evaluated on a dataset of 1K of Egyptian dialect sentences (including training and test 800 and 200, respectively). For converting dialect words to MSA words, the system achieved an accuracy of 88%, whereas for producing these words into their correct order the system performed 78%.

Elissa (Salloum and Habash, 2012) is a rule-based machine translation system from Arabic dialects to MSA. It handles Levantine, Egyptian, Iraqi, and to a lesser degree Gulf Arabic. After identifying dialectal words in a source sentence, Elissa produces MSA paraphrases using ADAM (Salloum and Habash, 2011) dialectal morphological analyzer, morphological transfer rules and dialect-MSA dictionaries. These paraphrases are used to form an MSA lattice that passes through a language model (LM) for n-best decoding and then selects the best MSA translations. In this paper, no evaluation has been provided.[3]

Mohamed et al. (2012) presented a rule-based approach to produce Colloquial Egyptian Arabic (CEA) from modern standard Arabic, they provide an application case to the Part-Of-Speech (POS) tagging task for which the accuracy has been improved from 73.24% to 86.84% on unseen CEA text, and the percentage of Out-Of-Vocabulary (OOV) words decreased from 28.98% to 16.66%.

Al-Gaphari and Al-Yadoumi (2012) used a rule-based approach to convert Sanaani dialect to MSA. Their system reached 77.32% of accuracy when tested on a Sanani corpus of 9386 words.

[3]Elissa is evaluated later in (Salloum and Habash, 2013).

Hamdi et al. (2013) presented a translation system between MSA and Tunisian dialect verbal forms. The work is based on deep morphological representations of roots and patterns which is an important feature of Arabic and its variants (dialects). The approach is similar to that used in (Mohamed et al., 2012), (Sawaf, 2010) and (Salloum and Habash, 2013) but is characterized by a deep morphological representation based on MAGEAD (Habash and Rambow, 2006) (morphological analyzer and generator for the Arabic dialects). The system translates in both directions (MSA to Tunisian dialect and vice versa). It reached a recall of 84% from dialect to MSA and 80% in the opposite side.

For translating Moroccan dialect to MSA, a rule-based approach relying on a language model was used in (Tachicart and Bouzoubaa, 2014). The system is based on a morphological analysis with Alkhalil morphological analyzer (Boudlal et al., 2010) adapted and extended with Moroccan dialects affixes and a bilingual dictionary (built from television productions scenarios and data collected from the web). After an identification step which separates dialectal data from MSA, the text is analyzed and segmented into annotated dialect units. These outputs are linked into one or more MSA corresponding units by using the bilingual dictionary. In the generation step, MSA phrases are produced then passed to a language model to produce the most fluent MSA sentences (no evaluation was given for this work).

Sadat et al. (2014) provided a framework for translating Tunisian dialect text of social media into MSA. The work is based on a bilingual lexicon created for this context. It adopts a set of grammatical mapping rules with a disambiguation step which relies on a language modeling of MSA for the selection of the best translation phrases. It should be noted that the translation system is word-based. It performs a BLEU (Papineni et al., 2002) score of 14.32 on a test set of 50 Tunisian dialect sentences (the reference was made by hand).

Meftouh et al. (2015) presented PADIC a multi-dialect Arabic corpus that includes MSA, Maghrebi dialects (Algerian and Tunisian) and Levantine dialects (Palestinian and Syrian). Unlike other contributions, several experiments were performed on different SMT systems between all pairs of languages (MSA and dialects). The authors analyzed the impact of the language model on machine translation by varying the smoothing techniques and by interpolating it with a larger one. The best results of translation were achieved between the dialects of Algeria which is not a surprising result since they share a large part of the vocabulary. It was also shown that the

Table 1: MT work between Arabic dialects and MSA: (Source and Target languages)

Work	Source	Target
(Bakr et al., 2008)	Egyptian	MSA
(Salloum and Habash, 2012)	Levantine, Egyptian, Iraqi, Gulf Arabic	MSA
(Mohamed et al., 2012)	MSA	Egyptian
(Al-Gaphari and Al-Yadoumi, 2012)	Sanaani (Yemenite)	MSA
(Hamdi et al., 2013)	Tunisian	MSA
	MSA	Tunisian
(Tachicart and Bouzoubaa, 2014)	Moroccan	MSA
(Sadat et al., 2014)	Tunisian	MSA
(Meftouh et al., 2015)	Algerian, Tunisian, Syrian and Palestinian	MSA
	MSA	Algerian, Tunisian, Syrian and Palestinian

performance of machine translation between Palestinian and Syrian was relatively high because of the closeness of the two dialects. Concerning MSA, the best results of machine translation have been achieved with Palestinian dialect.

In Table 1, we summarize all the work cited above in terms of concerned dialects and translation direction.

3.2. Translating between Arabic dialects and foreign languages

Sawaf (2010) built a hybrid MT system combining statistical and rule-based approaches. This system translates from Arabic dialects (spontaneous and noisy text from broadcast transmissions and web content) to English using MSA as pivot language. Dialect texts were normalized into MSA using character-based rules which utilizes simple rules to convert words into the most similar MSA words, then the text is analyzed by a dialect-specific and a MSA morphological analyzers. The results are entered into dialect normalization decoder that relies on language models and a lexicon. The work deals with a set of Arabic dialects: Levantine (Lebanese, North Syria, Damascus, Palestine and Jordan), Gulf Arabic (Northern Iraq, Baghdad, Southern Iraq, Gulf, Saudi-Arabia, and Southern Arabic Peninsula), Nile Region (Egypt and Sudan) and Maghreb Arabic (Libya, Morocco and Tunisia). Achieved results showed that hybrid MT performs better than statistical MT and rule-based MT and that normalizing and processing the text (both training and test corpora) improve translation quality in terms of BLEU by 2% for Web text and about 1% for broadcast news/conversations.

In (Salloum and Habash, 2011), the authors improved an Arabic-English SMT system by producing MSA paraphrases for OOV dialectal words and low-frequency words through a light-weight rule-based approach. They created ADAM (Arabic Dialect Morphological Analyzer) by extending the well-known BAMA (Buckwalter, 2004) with Levantine/Egyptian dialectal affixes and clitics. In addition to ADAM, they used a set of hand-write morpho-syntactic transfer rules. This allows to generating paraphrases that are input as a lattice to a state-of-the-art phrase-based SMT system. This last point is the main difference between this work and the one presented above (Sawaf, 2010). The latter produces unique MSA version for a dialect word where the former produces multiple MSA paraphrases (or alternative normalizations). Two SMT systems were built within this work, they were trained on two different data conditions, a MSA(only)-English parallel corpus (of 12M words on the Arabic side) and a large (MSA&Dialect)–English parallel corpus (of 64M words on the Arabic side). When evaluated on a blind test set, the SMT system trained on the large corpus using ADAM and transfer rules outperformed the baseline system (SMT system trained on the same data) by 0.56 absolute BLEU.

The same authors in (Salloum and Habash, 2013), presented a manual evaluation of Elissa (cited above). It was shown that 93% of MSA sentences produced by Elissa were correct. In addition, Elissa was used for pivoting through MSA in a dialect-English SMT system whose BLEU score was improved between 0.6% and 1.4%.

Sajjad et al. (2013) provided a dialectal Egyptian Arabic to English statistical machine translation system. They converted Egyptian to MSA by applying a character level transformational model (including morphological, phonological and spelling changes) learned from Egyptian-MSA words pairs. The MT system built on the adapted parallel data showed improvement in the quality of machine translation. Transformation task reduces the OOV words rate from 5.2% to 2.6% and improves BLEU score by 1.87 points. Whereas adapting large MSA/English parallel data gives significant reduction of OOV rate to 0.7% and leads to an absolute BLEU increase of 2.73 points.

Salloum et al. (2014) explored the impact of sentence-level dialect identification used with various linguistic features on machine translation performance. They attempted to optimize the selection of outputs produced by different MT systems given an input text including a mixture of dialects and MSA. The study concerns machine translation from Arabic dialect, namely Egyptian and Levantine to English. Four MT systems were used for this purpose, the first three ones are SMT systems trained on different corpora[4]: dialect-English (5M tokenized words of Egyptian and Levantine), MSA-English (57M tokenized words) and dialect+MSA-English (62M tokenized words). The fourth one is a MSA-pivoting system that combines dialect-to-MSA MT system (Salloum and Habash, 2013) and an Arabic-English SMT system. This last system is trained on dialect+MSA-English corpus augmented with dialect-English corpus where the dialectal side has been preprocessed with the dialect-MSA MT previously cited (Salloum and Habash, 2013). The size of this training corpus is 67M. We note that the MSA-pivoting system (the fourth one) produces the best BLEU score among all systems, it is the first baseline system. In this wrok, the same MT algorithms are used for training, tuning and testing each MT system, but as regards data each system is trained on a different dataset (as we saw above) in terms of the degree of source language dialectness. An interesting approach was adopted in this reaserach, instead of finding the most performant MT, the authors tried to identify automatically the most suitable MT system for a given sentence. They assume that these systems complement each other and combining their selections could lead to better overall performance. A baseline MT system selection based on a binary classification was built by using a sentence-level dialect identifier Elfardy and Diab (2013). This baseline selection system decides what MT system to use among the four systems described above. According to the authors, the best configuration defined is to select the MSA-English system for sentences tagged as MSA sentences and MSA-pivoting for sentences tagged as dialectal ones. The main contribution of this work is a MT selection system created using machine learning techniques trained on only source language features to select the best MT system that should translate each sentence in the test set. This selection system is a Naive Bayes Classifier (NBC) with four classes corresponding to the four MT systems. The training data of the classifier is a set of 5562 sentences labeled with the class label of the MT system that has produced the highest BLEU score (at sentence-level). The NBC uses a set of basic features such as: token-Level features which use language models, MSA & dialectal morphological analyzers and a dialectal lexicon (to decide whether

[4]Similar to (Zbib et al., 2012) discussed further below.

each word is MSA, dialectal, both, or OOV), perplexity features that include two features related to the perplexity of a sentence computed on the two languages models (MSA and dialect). In addition, the classifier uses some extended features extracted from the cited dialect-MSA MT system like sentence length (in words), number of punctuation marks, and number of words that are written in Latin script. Another set of extended features are used like the sentence perplexity computed on each source-side of the training data of each of the four MT systems. Using the NBC to predict the best MT system to use for translating a sentence had improved the BLEU score by 1% over the best score recorded for a single MT (which corresponds to the MSA-pivoting system). It also outperforms the baseline selection system by 0.6% BLEU.

Jeblee et al. (2014) presented a SMT system that translates (in contrast to all other research efforts) from English to Arabic dialect by pivoting through MSA. The translator is based on a core SMT system trained on a parallel English-MSA corpus of (5M pairs of sentences), the output of this system is translated to Egyptian dialect by using both dialect and domain adaptation system. It should be noted that for adaptation systems the authors created a tri-side parallel corpus (English, MSA and Egyptian dialect) of 100k sentences by using a rule-based method. For convenience of reading we refer to each side of this corpus as Eng-100k, MSA-100k and Egy-100k. Two variants of adaptation system were presented. The first variant translated with the core SMT system the English side (Eng-100k) of the tri-parallel corpus to MSA (we call the result MSA-100k-trans). This dataset is used with the Egyptian side (Egy-100k) of the corpus as training data to translate from MSA to Egyptian. An English test set is translated to MSA (by using the core SMT English-MSA), the result is then translated to Egyptian dialect by using the SMT trained on the parallel corpus (MSA-100k-trans, Egy100k). The second variant includes two adaptation steps. The first one is used to adapt the MSA output of the core system to the domain of the MSA side in the tri-parallel corpus and a second one to translate the MSA output of the domain adaptation system into Egyptian Arabic. An English test set is translated to MSA with the SMT core system, the result is then translated by the first adaptation system trained on (MSA-100k-trans, MSA-100k). The output of this step is then translated into Egyptian by using the second adaptation system trained on (MSA100k, Egy100k). The main result of this work showed that it is possible to increase the MT quality by using domain adaptation between MSA and Egyptian dialect as adapting between different domains of the same language. Furthermore, using MSA as a pivot then adapting to dialect could improve MT performance.

Al-Mannai et al. (2014) proposed an unsupervised morphological segmentation for Arabic dialects to improve machine translation quality. The study concerned a Qatari Arabic to English SMT. It was shown that segmentation with Morfessor (Siivola et al., 2007) (unsupervised morphological segmenter) improves the translation quality compared to a system without segmentation at all or to a system using Arabic Treebank (ATB) segmentation. In addition, a multi-dialectal word segmentation model was trained on the Arabic part of a parallel corpus including Qatari Arabic, Egyptian, Levantine, MSA and English. This segmented corpus was used to train the Qatari Arabic to English SMT, the BLEU score increased by 1.5 points when compared to a baseline system which does not use segmentation. In the other direction, a preliminary SMT system was trained to translate English to Qatari Arabic using the same parallel corpus without segmentation and by training the language model with other dialect corpora. The best system shows an absolute improvement of 0.22 in terms of BLEU compared to the baseline system that only uses the Arabic side of the Qatari Arabic corpus for language model (LM) training.

Durrani et al. (2014) improved Egyptian-to-English translation quality by handling OOV words. They first proceed to convert Egyptian to MSA by using a large monolingual language model to score the MSA-candidates for Egyptian OOV words (via a stack-based search with a beam-search algorithm). These candidates are got mainly through spelling correction and suggesting synonyms on context, MSA results are then translated to English via a SMT system. They showed that the spelling-based correction could improve the BLEU score by 1.7 points over the baseline system that translates unedited Egyptian into English. This work introduced an interesting idea to map Egyptian words into MSA by applying a convolution model using English as a pivot, the model relies on two corpora of 8.5K parallel sentences of Egyptian-English and 300K sentences of MSA-English.

Bolt Project[5]. DARPA launched the Broad Operational Language Translation (BOLT) program (2011-2014) to attempt to create new techniques for automated translation and linguistic analysis that can be applied to the informal genres of text and speech common in online and in-person communication

[5]http://www.darpa.mil/program/broad-operational-language-translation

in English, Chinese and Egyptian Arabic. BOLT has three technical areas: developing algorithms and integrated systems to support the translation, data collection and an evaluation step. Under this program, in (Zbib et al., 2012), two parallel corpora Levantine-English (1.1M words) and Egyptian-English (380K words) were built by translating parts extracted from a large corpus of Arabic web texts to English. Classification by dialect and translation were done by using Amazon's Mechanical Turk. Authors performed several experiments on a SMT system using these corpora in addition to a MSA-English parallel corpus (150M tokens for Arabic side). It was shown that morphological segmentation (using MADA (Habash and Rambow, 2005) morphological analyzer) uniformly improves translation quality. The work studied also the impact of dialectal training data size on MT performance. They show that a system trained on the combined dialectal-MSA data is likely to give the best performance, since informal Arabic data is usually a mixture of dialectal Arabic and MSA. Another interesting result was presented regards to pivoting through MSA or translating directly from dialect into English (the experiment was performed for Levantine only). The performance of the system improves by 2.3 BLEU points when pivoting through MSA for first experiment, but when adding more dialectal data to training set (400k words) direct translation becomes better than mapping to MSA despite the significantly low OOV rate with MSA-mapping.

Aminian et al. (2014) dealt with OOV words in the context of Arabic to English SMT system. They adopted an approach that normalizes dialectal words to MSA words by using AIDA[6](Elfardy et al., 2014) and MADAMIRA[7](Pasha et al., 2014), to identify and replace dialectal Arabic OOV words with their MSA equivalents. When tested on a blind dataset test, this approach improved SMT quality by 0.4% and 0.3% absolute BLEU for AIDA and MADAMIRA, respectively.

Within the same program, in (Aransa, 2015) a focus was made on Arabic dialect to English translation especially for Egyptian dialect. Several techniques have been implemented such as adapting SMT systems to the Egyptian dialect since the available training corpora, in the context of Bolt project, contain MSA and several dialects (Egyptian, Levantine and Iraqi).

[6]A dialect identification tool that identifies and classifies dialectal words on the token and sentence levels.

[7]A morphological analysis and disambiguation system for MSA and Egyptian dialect.

The performance of the system were improved by considering and treating the different dialects as different domains. An example of adaptation technique is using instance weighting of translation models to improve the translation quality by giving more weights to Egyptian than MSA or other Arabic dialects. It should be noted that the systems were adapted by using data selection techniques because the training data include various genres (News, Web, Discussion forums, SMS/CHAT). Data selection techniques consist of selecting the relevant sentences from monolingual corpora to improve and adapt the language models, or selecting the most relevant sentences from the bilingual corpora to improve the translation models. Another possible way of improving the system performance and translation quality was morphological segmentation. Several segmentation schemes were evaluated. Furthermore, in order to deal with the out-of-vocabulary words and to decrease the OOV rate proper noun transliteration was performed.

Recently, as regards the script used in dialectal texts, a new research line has been open up for Arabic dialect MT. It concerns Arabizi, also known as Romanized Arabic or Arabish. Arabizi is a non-standard writing system that uses Latin characters[8] to write Arabic dialects. It is widely used in the context of social media communications like Facebook, Twitter and YouTube, chat rooms and SMS. Arabizi is a mixture of both transliteration and transcription mappings, it does not obey to strict rules, it differs from one dialect to another, even in the same dialect community it differs from one user to another. Despite it has no standard form, a large amount of Arabizi data is generated by everyday communication (social media, SMS, etc). Thus, Arabizi creates new needs in the area of dialect NLP, it brings new challenges, especially for Machine Translation. It should be noted that the NIST OpenMT15[9] evaluation competition focused on informal data genres (SMS/Chat and Conversational Telephone Speech (CTS)) in Arabic dialect, precisely Egyptian, and Mandarin Chinese.[10] The task consisted in translating from Egyptian dialect and Mandarin Chinese into English.[11] It is worth noting that Egyptian dialect data within this campaign is a mixture of texts

[8] Including letters and numbers

[9] Open Machine Translation 2015

[10] `https://www.nist.gov/sites/default/files/documents/itl/iad/mig/OpenMT15_EvalPlan_v0-9.pdf`

[11] Official Evaluation results of NIST openMT15 are available in `ftp://jaguar.ncsl.nist.gov/mt/mt2015/openmt15results.html`

Table 2: MT work between Arabic dialects and English: Source/Target and MSA pivoting

Work	Source	MSA Pivoting	Target
(Sawaf, 2010)	Levantine, Gulf Arabic, Egyptian, Sudanese, Libyan, Moroccan, Tunisian	Yes	English
(Salloum and Habash, 2011)	Levantine, Egyptian	Yes	English
(Zbib et al., 2012)	Levantine, Egyptian	No	English
(Salloum and Habash, 2013)	Levantine, Egyptian, Iraqi, Gulf Arabic	Yes	English
(Sajjad et al., 2013)	Egyptian	Yes	English
(Jeblee et al., 2014)	English	Yes	Egyptian
(Al-Mannai et al., 2014)	Qatari	No	English
(Durrani et al., 2014)	Egyptian	Yes	English
(Aminian et al., 2014)	Egyptian	Yes	English
(Salloum et al., 2014)	Levantine, Egyptian	Yes	English
(May et al., 2014)	Egyptian	No	English
(Aransa, 2015)	Egyptian	No	English
(Van der Wees et al., 2016)	Egyptian	No	English

in both Arabic script and Arabizi.

In this respect, May et al. (2014) presented a SMT system that translates informal Egyptian dialect to English which deals with Arabizi. In this study, the authors created a deromanization module (converts Arabizi to Arabic script) whose output is translated into English via a SMT system trained on informal Arabic/English parallel and monolingual data (from DARPA BOLT). Their deromanization approach uses a character-based weighted finite state transducers (wFSTs) Mohri (1997) with a 5-gram character-based language model of Arabic dialect (learned from 5.4M words). We note that a character-based language model is used instead of a word-based one to avoid OOV words. Three methods were experimented to build Arabizi-to-Arabic script wFST, (1) manually by human experts[12], (2) automatically by using machine translation and (3) hybrid method (combining the two last). The first method consists in asking a native Arabic speaker to generate probabilistic character sequence pairs in order to encode the wFST transitions, whereas the automatic method is a SMT system trained on a corpus of 863 Arabizi/Arabic dialect (Arabic script) word pairs (where the words pairs are

[12]Familiar with finite-state machines

treated as sentence pairs and character are treated as words). According to the authors, this method produces more correspondences than the manual method and sequence pairs with longer context but generates also a set of noisy pairs that are useless. Another negative aspect of this method is that it does not generate vowel-dropping sequence pairs (that are taken into account by the first method). The hybrid method involves using sequences pairs (with Arabizi length of less than three characters) from those generated by the SMT system in addition to vowel-dropping sequence pairs from the manual wFST pairs. For the evaluation of both the deromanization module and the Arabizi-English SMT, the authors used two parallel corpora of Arabizi-English of 7,794 and 27,901 aligned sentences with reference deromanizations of the Arabizi side of each corpus. For the deromanization module, the automatic and hybrid methods outperform the manual one. However, the results of the hybrid approach are slightly better than the automatic approach. As regards the SMT systems scores, they track those of deromanization results. The SMT using automatically learned wFST approach outperforms the manual wFST (BLEU scores of respectively 12.0 and 8.9 Vs 15.1 and 13.2). In addition, the BLEU score (15.3 and 13.4) of the SMT system using the hybrid approach outperforms slightly the score of the SMT system using the automatic approach (15.1 and 13.2).

Van der Wees et al. (2016) attempted to improve Arabizi-to-English machine translation by using an Arabizi-to-Arabic script converter that does not require human knowledge (experts or native Arabic speakers). This converter has been incorporated into a phrase-based SMT system whose performance yields results that are comparable to those achieved after human transliteration. This work uses a set of resources including : a large Arabic dialect-English parallel corpus (1.75M sentence pairs with 52.9M Arabic tokens), a small tri-text Arabizi-dialect (Arabic script)-English (10K parallel sentences belonging to the SMS and chat genres[13]) from which 1788 parallel Arabizi-dialect (Arabic script) sentences were split into two test sets for evaluation, and finally, an Arabizi-English parallel corpus[14] crawled from a variety of web pages (10K sentence pairs with 180K Arabizi tokens). The first step of deromanization is generating transliteration candidates, this is done

[13]LDC catalog number: LDC2013E125, data set released for the most recent NIST OpenMT

[14]This resource has been created in the context of this work but the authors did not give any details about how they proceed.

by character mapping module [15] following the phrase-based SMT paradigm. Since the generated candidates could include character sequences that are not actual Arabic words, they are filtered by comparing them to a large Arabic dialect vocabulary (200K of distinct words) and the OOV candidates are then eliminated. This operation reduced the number of candidates for a given Arabizi word by 50% and also excluded Arabizi words with character repetitions.[16] After generating candidates and filtering steps, an ambiguous Arabizi-to-dialect(Arabic script) lexicon is created. This lexicon, in addition to a 3-gram Arabic dialect language model (trained on the source side of the available parallel dialect (Arabic script)-English corpora) are passed through a contextual disambiguation process using srilm-disambig[17] in order to search for the best transliteration of each Arabizi sentence. At this stage (we call it a first variant of the romanization), the WERs (Word Error Rates) recorded for the two set tests were 46.4% and 50.8%. For improving these results, the authors exploited transliterated word pairs extracted from the tritext Arabizi-dialect (Arabic script)-English described above. They added them to the transliterated lexicon used by the contextual disambiguation by prioritizing them with a high score (0.9 Vs 0.1 for the other transliteration candidates). This step (the second variant of deromanization) contributed to an improvement of the WERs by 50% (25.7% and 027.9%) on the two test sets. This transliteration module has been incorporated into an in-house phrase-based SMT trained on the collection of dialect (Arabic script)-English corpora described above (1.75M parallel sentences with 52.9M Arabic tokens) and a 5-gram English language model. On the other hand, the Arabizi-English corpus of web-crawled user comments has been used to train a small SMT system whose phrase translation and phrase reordering models have been merged to the main SMT system models. This increases the chance of translating (directly by the Arabizi-English models) a non-transliterated Arabizi word. For the two variants of the transliteration module, the SMT system has been evaluated using BLEU score. The best BLEU is recorded for the transliteration module that uses character-level mapping with contextual disambiguation augmented by words pairs (second variant) with 8.68

[15]The mapping of Arabic letters to Arabizi character sequences uses the publicly available character table described in http://en.wikipedia.org/wiki/Arabic_chat_alphabet

[16]character repetition is widely used in social media networks, SMS and Chat in order to lay emphasis on the word where it (the repetition) appears).

[17]http://www.speech.sri.com/projects/srilm/manpages/disambig.1.html

and 10.32 on the two test sets Vs a BLEU of 7.46 and 9.42 (for the first variant). Table 2 provides a summary of MT work listed above with regard to concerned dialects, translation direction and pivoting through MSA.

MuDMaT. Another project dedicated to machine translation of Arabic dialects is MuDMAT project (Multi-Dialect Machine Translation) (Sadat, 2015) supported by NSERC.[18] MuDMaT is speared over the period of (2014-2017). It aims to build MT systems between Maghreb dialects (Algerian, Moroccan and Tunisian), MSA and French using hybrid approach. According to the author, a demonstration of a rule-based machine translation from Tunisian dialect to MSA and French was achieved.

All the work cited above is related to text machine translation. For speech translation there are no relevant projects dedicated for Arabic dialects, except those funded by DARPA such as TRANSTAC [19] project (Hsiao et al., 2006), a predecessor program to BOLT which deals with MT between Iraqi dialect and English. The goal of TRANSTAC is a rapid development of bi-directional translation systems that allow speakers of different languages to communicate in real-world tactical situations. Several prototype systems were developed for military and medical screening domains to enable conversations with local foreign language speakers of Iraqi Arabic, Mandarin, Farsi, Pashto, and Thai. Some research was dedicated to evaluate MT scores of Iraqi Arabic and English translators such as (Condon et al., 2010) and (Condon et al., 2008). In the same context, IBM MASTOR (Gao et al., 2006), is a speech-to-speech translation system that translates spontaneous free-form speech in real-time on both laptop and hand-held PDAs for two language pairs, English-Mandarin Chinese, and English-Arabic dialect.

4. Discussion

We presented above a set of recent machine translation studies dedicated to Arabic dialects. This research work has been described in terms of used approach, data configuration and relevant results. In the following, we sum up the most significant findings of these different contributions:

- The limited number of covered languages shows that MT for Arabic dialects is just beginning. Indeed, all contributions are dedicated to

[18]National Science and Engineering Research Council of Canada.
[19]The Spoken Language Communication and Translation System for Tactical Use

translate between dialects, MSA and English. We note that there is only one work which translates to French but unfortunately, no results are available for it. In terms of translation direction, most of the contributions translate from dialects to MSA or English, whereas there is very little work that uses the dialect as target language. This may be explained by the fact that using dialect as target language for a SMT system for example requires important amount of cleaned data in order to build reliable language models. Even for rule-based MT systems, it requires adapted tools (morphological, syntactic and semantic generators). Such requirements are still unavailable for most Arabic dialects.

- Regards to the used dialects (see Table 3), it is clear that middle-east dialects are the most used ones especially Egyptian (spoken in the most populous Arab country[20]), followed by Levantine, whilst Maghrebi dialects are less present when for the other dialects like Koweitian, Bahraini, Omani and Mauritanian no work in this field was found.

[20]The current population of Egypt is 94,899,254, based on the latest United Nations estimates.

Table 3: Arabic dialects concerned by MT research

Dialect	Translation Work between	
	MSA and dialects	dialects and English
Egyptian	(Bakr et al., 2008), (Salloum and Habash, 2012), (Mohamed et al., 2012)	(Sawaf, 2010), (Zbib et al., 2012), (Salloum and Habash, 2013), (Sajjad et al., 2013), (Jeblee et al., 2014), (Aminian et al., 2014), (Durrani et al., 2014), (Salloum et al., 2014), (May et al., 2014), (Aransa, 2015), (Van der Wees et al., 2016)
Levantine	(Salloum and Habash, 2012), (Meftouh et al., 2015)	(Sawaf, 2010), (Zbib et al., 2012), (Salloum and Habash, 2013), (Salloum et al., 2014)
Tunisian	(Hamdi et al., 2013), (Sadat et al., 2014), (Meftouh et al., 2015)	(Sawaf, 2010)
Iraqi	(Salloum and Habash, 2012)	(Sawaf, 2010), (Salloum and Habash, 2013)
Gulf Arabic	(Salloum and Habash, 2012)	(Sawaf, 2010), (Salloum and Habash, 2013)
Moroccan	(Tachicart and Bouzoubaa, 2014)	(Sawaf, 2010)
Sanaani (Yemenite)	(Al-Gaphari and Al-Yadoumi, 2012)	
Algerian	(Meftouh et al., 2015)	
Sudanese		(Sawaf, 2010)
Libyan		(Sawaf, 2010)
Qatari		(Al-Mannai et al., 2014)

- In terms of methodology, for translating between MSA and dialects the rule based approach with morphological analysis is the most used method. In addition, most work exploit bilingual lexicons and rely on relatively small language models (see data description in Table 4) compared to those used for standard languages. We can see that Egyptian and Levantine to a lesser degree, are in advance compared to other dialects. Recent work on Moroccan, Tunisian and Yemenite dialects adopt almost the same approach that was used in the first studies of Egyptian and Levantine. It is clear that when no relevant corpora are available, the rule-based approach is adopted despite its drawbacks.

Table 4: MT work between Dialects and MSA: Approaches, data description and results

Work & Best results	Approach	Data description
(Bakr et al., 2008) Accuracy: 88%	Statistical tokenization & tagging + Rule-based transformation	1k sentences
(Salloum and Habash, 2012) Accuracy: 93.15%	Rule-based approach + LM	300 sentences
(Mohamed et al., 2012) POS tagging evaluation Accuracy: 73.24%	Rule-based approach	100 user comments
(Al-Gaphari and Al-Yadoumi, 2012) Accuracy: 77.32%	Rule-based approach	9386 words
(Hamdi et al., 2013) Accuracy: Tunisian-to-MSA 84% MSA-to-Tunisian 80%	Rule based approach (deep morphological representation of data)	Parallel Tunisian/MSA corpus of 1500 sentence pairs Dev/test set 750 sentence pairs,
(Sadat et al., 2014) BLEU score: 14.32	Rule-based approach +Bilingual lexicon+LM	50 sentences
(Tachicart and Bouzoubaa, 2014)	Rule-based approach +Bilingual lexicon+LM	-
(Meftouh et al., 2015) A set of BLEU scores	Statistical approach	6 sides parallel corpus of 6400 sentences Dev/test set 500 sentence for each corpus

- For machine translation between Arabic dialects and English, the dominant methodology is hybridizing rule-based and statistical approaches, especially for the first research work (see table 5). The SMT systems are trained on large MSA/English corpora in addition to relatively smaller dialectal corpora. The rule-based methods rely on morphological analysis and transfer rules to normalize dialectal words into MSA words (Sawaf, 2010; Salloum and Habash, 2011, 2013; Sajjad et al., 2013). Other work uses domain adaptation techniques by considering dialect adaptation as a domain adaption problem (Sajjad et al., 2013; Jeblee et al., 2014; Al-Mannai et al., 2014; Aransa, 2015). The availability of some parallel corpora makes this research direction possible. Furthermore, availability of new tools related to dialect identification (at word and sentences levels) has a positive impact on machine translation performance as it was shown in (Aminian et al., 2014; Salloum et al., 2014). Indeed, in this last work, identifying either the sentence is dialectal or MSA guides the selection of the MT system to use. Also, It must be stressed that the impact of segmentation have been showed in most work, it improves MT scores significantly.

- An important point related to Arabic dialects MT is using MSA as a pivot language when translating to or from English. As mentionned above, exploiting the proximity between close languages has been used in NLP research dedicated to under-resourced languages. The idea is to adapt existing resources of a rich-resourced language to process an under-resourced language, particularly, in the context of standard langauges and their dialects. This research direction has been adopted in the area of Arabic dialects NLP and especially for machine translation. We observe that the first efforts were (naturally) dedicated to translating between dialects and MSA, probably with a view to reaching other standard languages. Pivoting through MSA has been used in a majority of contributions, they state that it improves MT quality, except one work (Zbib et al., 2012) which shows that increasing the dialect training data increases MT performance better than pivoting through MSA, but it noticed that the OOV rate is lower with MSA-mapping. The authors concluded that differences in genre between MSA and dialects make vocabulary coverage insufficient and considering the domain is an important research direction. We note that an intersting idea has been introduced in the work of Salloum et al. (2014) where authors have combined four MT systems among them a system which pivots through MSA and a system that translates directly from dialect to English. By using learning machine techniques and according to dialect level of the sentence, they select the adequate MT system (among the four ones) to translate (the considered sentence). Consequently, they continue to take profit of MSA-mapping whenever it is possible. In this way, the MT systems form a whole and complement each other.

- Another big challenge of Arabic dialects MT is Arabizi (aka Arabish or Romanized Arabic). Indeed, important amount of user-generated data from social networks are a mixture of dialect written in both Arabic and Roman script. Given their size, these data could be an important source of dialect corpora if they are processed. It is in this context that most recent MT work for Arabic dialects attempt to deal with translation from Arabizi to English. But, despite its large use, Arabizi is still a new research direction, few work are dedicated to it and they concern only Egyptian dialect. Other Arabic dialects are at a preliminary stage. The contributions presented in this survey related to Arabizi are based on a SMT system built on the top of a deromanization module that converts Arabizi texts to Arabic script. The importance of deromanization is evident, it was shown that its accuracy rates correlate with MT scores (in both of the two papers presented above). We expect that future work attempt to bypass the step of deromanization when more parallel corpora including Arabizi will be available. Thus, Arabic-script pivoting and direct translation will be certainly experimented. In this respect, direct translation from Arabizi to English or French will probably reduces the complexity of two serious problems related to Arabic dialects MT; proper nouns translation and code-switching. Since Arabizi uses Roman script, there is no need to translate proper nouns even more English or French words[21] (included in dialect text in the case of code-switching)

- Regards to the data, we notice a significant lack of textual resources dedicated for dialects. All research efforts deal with this issue. We can see that MSA-dialect parallel corpora are fewer than English-dialect ones (see data description in Tables 4 and 5). This is due to the fact that MT projects between Arabic dialects and English are more funded mainly in the case of BOLT project. Yet, even with this funding, the corpora including dialects are smaller than those of standard languages (MSA/English for example). Also, in terms of coverage, Egyptian and Levantine remain the most resourced dialects in contrast to all others. It is worth noting that an important portion of several MT efforts is dedicated to produce dialect resources.

[21] In the Middle-east, Arab people switch between dialect, MSA and English, whereas in the Maghreb the code-switching is observed between dialect, MSA and French.

Table 5: MT work between Dialects and English: Approaches, data description and results

Work & Best results (BLEU)	Approach	Data description
(Sawaf, 2010) Broadcast News 36.4 Web content 42.1	SMT +Rule-based approach	Training/test(Dialect/English): Broadcast News 14.3M/ 12.4K sentences Web content 38.5K/ 547 sentences
(Salloum and Habash, 2011) 37.8	SMT +Rule-based approach	Training(MSA/English): 32M words (MSA side)(LDC2007E103) (Dialect&MSA-English)64M words (MSA side) dev&test sets of 1496&1568 sentences
(Zbib et al., 2012) Egyptian 20.66 Levantine 19.29	SMT + Morpho. segmentation	Training(dialect/English): 180k sentence pairs (1.1M Levantine ,380k Egyptian, English 2.3M words) Training(MSA/English):8M MSA-English sentence pairs
(Salloum and Habash, 2013) Dev10 set: 39.13 Levantine test set: 10.54 Egyptian test set: 19.59	SMT +Rule-based approach	Training(MSA/English): 64M words (MSA side) Dev10 test set 1568 sentences(audio dev data DARPA GALE program Levantine test set 2728 sentences (Zbib et al., 2012), Egyptian test set 1553 sentences (BOLT program)
(Sajjad et al., 2013) 16.96	Character-level transformational model +SMT +Data adaptation	Dialect/English parallel corpus of 38k sentences (Zbib et al., 2012) Training: 32k sentences ,Test: 4k sentences Training(MSA/English): 200k sentences from (LDC2004T17,LDC2004E72, & parallel corpora of the GALE program)
(Jeblee et al., 2014) 42.9	SMT +Domain adaptation +Dialect adaptation	Training set:English/MSA 5M parallel sentences(NIST 2012) Test set 1313 (NIST MT09) A 100k artificial tri-parallel Egyptian-MSA-English corpus
(Al-Mannai et al., 2014) 15.2	SMT+segmentation +Adapting MSA and other dialects	Qatari Arabic/English corpus(Elmahdy et al., 2014) Training set: 12k sentences Test set: 1k sentences
(Durrani et al., 2014) 23.72	Egyptian-to-MSA decoder + MSA-to-English decoder	Gale-dev10 set and Bolt Egyptian (tahyyes dev set)
(Aminian et al., 2014) AIDA 25.9 MADAMIRA: 28.8	SMT +Dialectal words identification and replacement	Training set:MSA side 29M tokenized words and Dialect side 5M DA tokenized words Dialect test set (BOLT-arz-test) 1065 sentences(LDC2012E30),16177 tokenized words MSA test set(MT09-test) 1445 sentences (LDC2010T23)40858 tokenized words
(Salloum et al., 2014) 33.5	Sentence level dialect identification +SMT selection using Naive Bayes Classifier	Training set:Dialect/English parallel corpus of 5M tokenized words (BOLT) MSA/English parallel corpus of 57M tokenized words NBC Training 2562 sentences Dev set/ Test set of 1802 & 1804 sentences
(May et al., 2014) 15.3 13.4	Deromanization system +SMT	Training Deromanization system Dialect corpus of 5.4 M words & 863 Arabizi/Arabic words Dialect/English parallel corpus (BOLT) Two test sets of 7794 & 27901 sentences
(Aransa, 2015) A set of scores BLEU and (Ter - Bleu)/2 (Servan and Schwenk, 2011)	SMT+ Language & translation models adaptation +Segmentation schemes +Proper nouns transliteration	Different datasets of(training/dev/test): Discussion forum SMS/Chat system Conversational telephone speech (CTS) transcript
(Van der Wees et al., 2016) 8.68 10.32	Deromanization system +SMT	Dialect/English parallel corpus of 1.75M sentence pairs Arabizi-Arabic-English corpus of 10K sentences (LDC2013E125) Arabizi-English corpus of 10K sentences(180K Arabizi tokens) 1788 pairs of sentences split into two test sets.

5. Conclusion

The above findings draw a picture of machine translation in the context of Arabic dialects. We can observe that dialects emerge as real languages and any NLP tools and resources dedicated to MSA should taking into account these dialects. Machine translation for Arabic dialects is still an immature area of research. There is still a long way to walk. Several important issues need to be solved. The dialects themselves, as they are presented in all the research work are classified by country or by region: Levantine dialect, Egyptian, Algerian, Tunisian, etc. This classification simplifies considerably the real linguistic situation through Arab countries. In fact, each Arab country has multiple varieties of dialects with specific features. MT systems dedicated to dialect have to deal with all these variants. In addition, the wide use of Arabizi in social networks generates new challenges that needs to be addressed also.

Another issue has to be taken into account is the code switching, Arab people switch in their conversation between dialect, Arabic and other languages, especially in the Maghreb where people tend to use French, Arabic, dialect and even Berber. This code-switching is a challenge for dialects MT. Also, it should be noted that for Maghreb dialect an important source of OOV words could be the use of French words, handling this issue must take into account this aspect since MSA pivoting or normalizing Maghreb dialect words to MSA could be insufficient. In the same vein, fast evolution of dialects needs to be considered for machine translation. Indeed, everyday new dialectal words appear and are adopted by people spontaneously without any official or academic validation.

As regards resources, a way to get parallel data is to use an iterative approach to produce artificial dialectal data from available dialect MT systems by post-editing their output. Another interesting track is to investigate comparable corpora for producing parallel corpora for training machine translation systems. This is already done for natural language such as in : (Jehl et al., 2012), (Hewavitharana and Vogel, 2011) for the pair Arabic-English, (Cettolo et al., 2010) for English-German and Arabic-English, (Munteanu and Marcu, 2006) for Romanian-English, and (Tillmann and Xu, 2009) for Spanish-English and Portuguese-English. This approach is feasible for Arabic dialects by using social networks which are a rich source containing a huge quantity of data expressed in dialects. But unfortunately, these noisy data require a considerable pre-processing steps such as: dialect identification, morphological analysis with specific tools, cleaning the data by eliminating non-exploitable fragments and writing normalization.

References

C. A. Ferguson, Diglossia, Word 15 (1959) 325–340.

N. Y. Habash, Introduction to Arabic natural language processing, Synthesis lectures on human language technologies 3 (1) (2010) 1–187, ISSN 1573-0573.

A. Shoufan, S. Al-Ameri, Natural Language Processing for Dialectical Arabic: A Survey, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistic (ACL), the Arabic Natural Language Processing workshop (ANLP), 36–48, 2015.

X. Zhang, Dialect MT: A Case Study between Cantonese and Mandarin, in: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING), Volume 2, Montreal, Quebec, Canada, 1460–1464, 1998.

J. Hajič, J. Hric, V. Kuboň, Machine translation of very close languages, in: Proceedings of the 6th Conference on Applied natural language processing (ANLC), Association for Computational Linguistics, 7–12, 2000.

K. Altintas, I. Cicekli, A machine translation system between a pair of closely related languages, in: Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS), 192–196, 2002.

K. P. Scannell, Machine translation for closely related language pairs, in: Proceedings of the Workshop Strategies for developing machine translation for minority languages, Citeseer, 103–109, 2006.

P. Nakov, H. T. Ng, Improving statistical machine translation for a resource-poor language using related resource-rich languages, Journal of Artificial Intelligence Research (4) (2012) 179–222.

B. Haddow, A. H. Huerta, F. Neubarth, H. Trost, Corpus development for machine translation between standard and dialectal varieties, in: Proceedings of Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, 7–14, 2013.

H. A. Bakr, K. Shaalan, I. Ziedan, A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic, in: Proceedings of the 6th International Conference on Informatics and Systems (INFOS). Cairo University, 2008.

W. Salloum, N. Habash, Elissa: A Dialectal to Standard Arabic Machine Translation System, in: 24th International Conference on Computational Linguistics (COLING), 385–392, 2012.

W. Salloum, N. Habash, Dialectal to Standard Arabic paraphrasing to improve Arabic-English statistical machine translation, in: Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, Association for Computational Linguistics, 10–21, 2011.

W. Salloum, N. Habash, Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies (HLT), 348–358, 2013.

E. Mohamed, B. Mohit, K. Oflazer, Transforming Standard Arabic to Colloquial Arabic, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic (ACL): Short Papers - Volume 2, 176–180, 2012.

G. Al-Gaphari, M. Al-Yadoumi, A method to convert Sanaani accent to Modern Standard Arabic, International Journal of Information Science and Management (IJISM) 8 (1) (2012) 39–49.

A. Hamdi, R. Boujelbane, N. Habash, A. Nasr, The effects of factorizing root and pattern mapping in bidirectional Tunisian-Standard Arabic machine translation, in: MT Summit, 2013.

H. Sawaf, Arabic dialect handling in hybrid machine translation, in: Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA), Denver, Colorado, 2010.

N. Habash, O. Rambow, MAGEAD: A morphological analyzer and generator for the Arabic dialects, in: Proceedings of the 21st International

Conference on Computational Linguistics (COLING) and the 44th annual meeting of the Association for Computational Linguistics (ACL), 681–688, 2006.

R. Tachicart, K. Bouzoubaa, A hybrid approach to translate Moroccan Arabic dialect, in: Proceedings of the 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14), IEEE, 1–5, 2014.

A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. O. A. o. Bebah, M. Shoul, Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts, in: Proceedings of the International Arab Conference on Information Technology, ACIT, 2010.

F. Sadat, F. Mallek, M. Boudabous, R. Sellami, A. Farzindar, Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Application, the case of Tunisian Arabic and the Social Media, in: Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, Association for Computational Linguistics and Dublin City University, 102–110, 2014.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 311–318, 2002.

K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, K. Smaili, Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus, in: Proceedings of the 29th Asia Conference on Language, Information and Computation (PACLIC), 26–34, 2015.

T. Buckwalter, Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, University of Pennsylvania, 2002. LDC Cat alog No.: LDC2004L02, Tech. Rep., ISBN 1-58563-324-0, 2004.

H. Sajjad, K. Darwish, Y. Belinkov, Translating Dialectal Arabic to English, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria, 1–6, 2013.

W. Salloum, H. Elfardy, L. Alamir-Salloum, N. Habash, M. Diab, Sentence Level Dialect Identification for Machine Translation System Selection, in:

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistic (ACL), 772–778, 2014.

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

H. Elfardy, M. T. Diab, Sentence Level Dialect Identification in Arabic, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic (ACL):, 456–461, 2013.

S. Jeblee, W. Feely, H. Bouamor, A. Lavie, N. Habash, K. Oflazer, Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Workshop on Arabic Natural Langauge Processing (ANLP), 196–206, 2014.

K. Al-Mannai, H. Sajjad, A. Khader, F. Al Obaidli, P. Nakov, S. Vogel, Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Workshop on Arabic Natural Langauge Processing (ANLP), 207–216, 2014.

V. Siivola, M. Creutz, M. Kurimo, Morfessor and variKN machine learning tools for speech and language technology, in: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 1549–1552, 2007.

N. Durrani, Y. Al-Onaizan, A. Ittycheriah, Improving Egyptian-to-English SMT by Mapping Egyptian into MSA, in: Proceedings of 15th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 271–282, 2014.

N. Habash, O. Rambow, 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 (ACL), 573–580, 2005.

M. Aminian, M. Ghoneim, M. Diab, Handling OOV Words in Dialectal Arabic to English Machine Translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Workshop Language Technology for Closely Related Languages and Language Variants (LT4CloseLang), 99–108, 2014.

H. Elfardy, M. Al-Badrashiny, M. Diab, AIDA: Identifying code switching in informal Arabic text, Proceedings of The First Workshop on Computational Approaches to Code Switching (2014) 94–101.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, R. M. Roth, Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic, in: Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.

W. Aransa, Statistical Machine Translation of the Arabic Dialect, Ph.D. thesis, University of Maine, doctoral school STIM, 2015.

J. May, Y. Benjira, A. Echihabi, An Arabizi-English social media statistical machine translation system, in: Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA), 329–341, 2014.

M. Van der Wees, A. Bisazza, C. Monz, A Simple but Effective Approach to Improve Arabizi-to-English Statistical Machine Translation, in: Proceedings of the International Conference on Computational Linguistics (COLING), Workshop on Noisy User-generated Text (WNUT), 43–50, 2016.

M. Mohri, Finite-State Transducers in language and speech processing, Computational linguistics 23 (2) (1997) 269–311.

F. Sadat, Multi-Dialect Machine Translation (MuDMat), in: Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT), Antalya, Turkey, 226, 2015.

R. Hsiao, A. Venugopal, T. Köhler, Y. Zhang, P. Charoenpornsawat, A. Zollmann, S. Vogel, A. W. Black, T. Schultz, A. Waibel, Optimizing components for handheld two-way speech translation for an English-Iraqi Arabic system., in: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech), 2006.

S. Condon, D. Parvaz, J. Aberdeen, C. Doran, A. Freeman, M. Awad, Evaluation of machine translation errors in English and Iraqi Arabic, Tech. Rep., 2010.

S. Condon, J. Phillips, C. Doran, J. Aberdeen, D. Parvaz, B. Oshika, G. Sanders, C. Schlenoff, Applying Automated Metrics to Speech Translation Dialogs, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC), 2008.

Y. Gao, L. Gu, B. Zhou, R. Sarikaya, M. Afify, H.-K. Kuo, W.-z. Zhu, Y. Deng, C. Prosser, W. Zhang, L. Besacier, IBM MASTOR System: Multilingual Automatic Speech-to-speech Translator, in: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies (HLT), the Workshop on Medical Speech Translation (MST), 57–60, 2006.

M. Elmahdy, M. Hasegawa-Johnson, E. Mustafawi, Development of a TV Broadcasts Speech Recognition System for Qatari Arabic, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC), Reyk-javik, Iceland, 2014.

C. Servan, H. Schwenk, Optimising multiple metrics with MERT, The Prague Bulletin of Mathematical Linguistics 96 (2011) 109–117.

L. Jehl, F. Hieber, S. Riezler, Twitter translation using translation-based cross-lingual retrieval, in: Proceedings of the 7th workshop on statistical machine translation, Association for Computational Linguistics, 410–421, 2012.

S. Hewavitharana, S. Vogel, Extracting Parallel Phrases from Comparable Data, in: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, Association for Computational Linguistics, Portland, Oregon, 61–68, 2011.

M. Cettolo, M. Federico, N. Bertoldi, Mining parallel fragments from comparable texts, in: Proceedings of the 7th International Workshop on Spoken Language Translation (IWSLT), 227–234, 2010.

D. S. Munteanu, D. Marcu, Extracting parallel sub-sentential fragments from non-parallel corpora, in: Proceedings of the 21st International Conference on Computational Linguistics (COLING) and the 44th annual meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 81–88, 2006.

C. Tillmann, J.-m. Xu, A simple sentence-level extraction algorithm for comparable data, in: Proceedings of the 2009 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies (HLT), Companion Volume: Short Papers, Association for Computational Linguistics, 93–96, 2009.