Developing a Transfer-Based System for Arabic Dialects Translation

Salwa Hamada and Reham M. Marzouk

Abstract The prominent Arabic Domestic changes have influenced the usage of the Arabic dialects among Arabs communications, which was, previously, limited on daily activities inside their own territories the role of the Modern Standard. Arabic MSA as an official Arabic language started to be diminished, since the Arabic dialects play a greater role than using it during the daily activities. The continuity of using these dialects whether in media or writing may eliminate the dominance of MSA as an official form of Arabic language in the Arab world. Besides, comprehending the Arabic language by non-native speakers, as well as, processing machine translations became a sophisticated process that requires harder effort. Accordingly, a requirement of language processing to interact with the permanent development of the dialects and to flourish the standard Arabic became imperative. Thus, it is planned to built a Hybrid Machine translation system (AlMoFseH) to translate the different Arabic dialects by using the MSA as a pivot. This research is a part of this project which emphasizes on developing a transfer-based system that transfers the Egyptian Arabic dialect EGY used in social media to MSA. For that purpose, a lexical database of 3k words presenting Egyptian Arabic dialect was built. Different texts extracted from Social media were used as a main resource of the database. The system consists of three components: disambiguation of the morphological analysis output using Naive Bayesian learning, a rule based transfer system and a dictionary look up system. The evaluation revealed a high accuracy of the system's performance, since 92.7% of the test data was transferred correctly.

Keywords Machine translation · Lexical database · Transfer system
Naive Bayesian algorithm · Egyptian Arabic dialect

S. Hamada (✉)
Electronics Research Institute ERI, P.O. Box 12611, Giza, Egypt
e-mail: hesalwa@hotmail.com

R.M. Marzouk
Faculty of Arts, Phonetics and Linguistics Department,
Alexandria University, P.O. Box 21526, Alexandria, Egypt
e-mail: marzoukreham@gmail.com

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_7

1 Introduction

Machine translation MT systems are computer programs that translate from a source language to a target language. The difficulties, that any machine translation confronts, are enlarged by the enlargement of the ambiguity in one or both languages. The Arabic language is one of these languages which the inflectional richness and sparsity of its dialects cause a large scale of ambiguities. Arabic language is classified as a diaglossic language where two forms of the language exist side by side; the formal form is known as Modern standard Language (MSA), while the other form is used in daily communication in each Arabic region and it is called dialect [1]. Both varieties form the linguistic repertoire of MSA written texts without clear boundaries between them [2]. Recently and Due to the permanent growing of the social media texts, Arabic dialects dominated these written texts and became an alternative of the formal MSA form. Thus, the recent studies of linguistics and language technology are directed to study these changes and their effect on the natural language processing. Besides, several sorts of texts were extracted and used as an essential resource for processing these dialects to unify their variations in one comprehensible form for machine translations and the non Arabic native speakers.

Otherwise, the social media user's inclination to improvise during writing enlarged the writing diversity, and added new ambiguities that should be taken into consideration. Accordingly, the need of developing a system that accepts these diversities during the translation of these texts became essential. Thus, developing hybrid system is planned to combine the best achievements of statistical and rules-based paradigms. This system is based on serial combinations of other multiple systems outputs.

This research describes two essential processes needed to achieve the system: the first process is building a lexical database to cover the words that occur frequently and signify the developed Egyptian dialect used in social media. The lexical database is stem based which is divided into lists provided with the required semantic and morpho-syntactic information of the dialectal stems and their equivalents in MSA. The second process describes the way this lexical database is incorporated into the translation process through a transfer system. The task of the transfer system is to normalize the social media texts that contain different varieties to reach to a standard MSA text that can consequently been translated to other language or dialect. This study will elaborate in details the process of transferring a nonstandard text into a standard one. This process requires incorporation of a statistical modeling for classification of the analyzed source text, normalization rules and functional rules to map the surface form of the analyzed texts into the closest form of the lexical database to facilitate the transferring.

Briefly, in this research:

A lexical framework was built to facilitate the selections of the lexical items whether as individual or multiple words. It covers the morphological and phonological distinctions between the Egyptian dialect, (which is taken as a representative dialect of this stage), and MSA in order to produce an underlying form of standard

Arabic sentences that can be generated in a further stage. The lexical items were selected carefully to cover the recent variations of the dialect according to a previous study of the Egyptian corpus ARZ ATB by Marzouk and El KareH (2016).

A transfer system was developed, with the enhancement of the lexical framework, to normalizes social media texts that are composed of a mixture of Arabic dialects and MSA.

Rewrite rules were created to approach the similarity of the morphosyntactic features for both sides, the source dialect and the target MSA.

The objective of the work is to unify the various written forms used in social texts in one standard form that can be comprehended and translated to other languages. The main contribution of this research is that the source of texts are different and significant from the usual texts previously used to present the written forms. Therefore, the results revealed the requirement of handling the new semantic and morphological ambiguities that caused by these texts. The paper is organized as following: Sect. 2 shades lights on the previous studies on dialects transfer and Arabic Machine translation, Sect. 3 describes the main issues that signify the Arabic language and its dialect, Sect. 4 overviews the main machine translation paradigms, Sect. 5 elicits the main modules that are involved in constructing the proposed system, Sect. 6 presents the procedures of building lexical database and the process of collecting the data, Sect. 7 is an evaluation of the systems performance and its results followed by the conclusion and the planned future work.

2 Related Studies

Previous apropos studies on dialect machine translation were limited and most of them were restricted on the normalization of one Arabic dialects words into their equivalents in MSA as a preliminary stage for their translation into English language [3]. Abo Bakr et al. (2008) explained the techniques of transferring Egyptian Arabic dialect into MSA and diacritization of the transferred MSA text. First they used a corpus collected from different pages from WWW to create the Egyptian colloquial to MSA lexicon. Then they depended on Buckwalter Arabic Morphological Analyzer BAMA to segment and analyze the source text. Support Vector Machine SVM multi-Classifier was used for the tokenization and diacritization. Moreover They added Segment type position to indicate the proper order of the segment in the target word or sentence, and new segment type position to move the segment to its proper order. The system's accuracy of converting Egyptian Arabic text into MSA showed that 88% were correct, and the diaritization of the MSA output's accuracy showed that 70% [4]. The main limitation of the work was the unavailability of a TreeBank that represents the Egyptian Arabic dialect. The collected corpus for this system represents a specific genre of the Egyptian Arabic text in social media which is a mixture of EGY and MSA, because written texts in these pages are directed to general communities and educated people. Therefore it may lack some linguistics forms that

signify the spoken dialect that transmitted to written texts. Moreover, Buckwalter morphological analyzer was originally designed in order to analyze MSA, therefore the analysis of dialectal data using Buckwalter analyzer in other studies was a cause of reduction in the output's accuracy.

A. Abdel Monem et al. (2008) investigated the usage of the interlingua machine translation approach for morphological and syntactic generation of Arabic texts. They followed rule based grammar generation approach to transform a semantic interlingual representation into Arabic texts. A. Abdel Monem et al. were the first who used the rule based approach from interlingua for morphological and syntactic generation of Arabic text [5]. For the evaluation they used English source sentences of approximately 1900 words and Arabic target sentences of 1600 words. The evaluation achieved a BLEU score with average 0.74, the results of the system performance was confirming the ability of the rule based approach to generate Arabic texts.

Sawaf (2010) developed a hybrid machine translation system to handle Arabic dialects by using a decoding algorithm that normalizes non-standard, spontaneous and dialectal Arabic into MSA. Sawaf's system goes through the following stages: Preprocessing and segmentation modules, Lexical Functional Grammar (LFG) system which incorporates a richly annotated lexicon containing functional and semantic information, functional models to use functional constrains to perform a deeper semantic and syntactic analysis for the source and target language, and Statistical translation models which use the maximum entropy framework [6]. The measured BLEU score of the system reached to 34.6%.

Salloum and Habash (2011) created ELISSA, the Dialectal Arabic DA to MSA Translation System. ELISSA is a ruled based model which relays on an existed morphological analyzer, DA-MSA dictionary and a model to rank and select the generated sentence. It follows certain steps to reach to its target: selection to identifying the word as a dialectal or out-of-vocabulary OOV, translation using classical rule based machine translation flow, morphological analysis ADAM, morphological transfer and morphological generation, and Language Modeling using SRILM for n-best decoding [7].

Other work that concentrated on the Arabic dialects translation, in specific Egyptian and Levantine Arabic was of Zbib et al. (2012), who developed a parallel corpus for the mentioned Arabic dialects using Crowdsourcing. Then they used the data in variant MT experiments. The parallel corpus which consists of 1.5M were classified according the dialects. The resulted dialects were attributed to 4 regions Levantines, Gulf area, Egypt, and Morocco, in addition to MSA. In the next step, The Levantine and Egyptian Arabic texts were translated by non professional translators using Amazon mechanical Turk. Zabib et al. performed a set of experiments to compare system trained using their parallel corpora with other systems trained on larger MSA-English parallel corpora. The experiments objectives were to investigate, first, the effect of the training data size, by examining different sizes of the training set, second, the cross dialect training, by using a training test of one dialect for a translate system of another, third, the validation of independent test data by using test set selected randomly from social media. Zabib et al. concluded that the 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. Also the mismatch between the training and the test data is the main reason beyond the lack of vocabulary coverage [3].

3 Arabic Language Variation

Arabic language is the forth widely spoken language [5]. More than 300 Millions people speak Arabic language [1]. MSA is descended from the Classical Arabic, the language of the Islamic holy book, "Quran". The syntax of MSA is unchanged from the classical Arabic but the changes affect its vocabulary and phraseology [8]. Nowadays, MSA is the written language of Arabic literature, journalism that stands side by side with the spoken regional vernaculars which are known as colloquial Arabic or Arabic dialects [9]. All native speakers learn their dialects as their mother tongue before they begin formal education [8]. These Arabic dialects are distributed along the Arab world from Morrocco in the west to Amman in the east [8]. Each country has its regional dialects but the mentioned dialects in these research are the capital cities dialects for their wide spreading comparing with the other regional dialects. The study of the spoken Arabic language has been dominated by the study of these dialects, but these studies were mostly confronted with negative attitude, as there is a worry that the study of a certain Arabic dialect may affect the supremacy of the study of the Standard Arabic [10]. Although, the Arabic dialects intervention and usage in a wide range of written texts couldn't be resisted for many reasons such as literature purposes in which some novels that talk about certain social and cultural level were preferred to be written using the slang language. As well as the spreading of electronic texts such as in SMS, chatting, and other communication media which became rich sources for the dialects in its written form. The Egyptian Arabic Dialect, and in specific the Cairene (the spoken colloquial Arabic of Egypts capital and the central Delta) is often considered to be the most widely understood dialect throughout the Arabic world [11]. This wide spread intelligibility is a result of the dominance of the Egyptian media in the Arabic world. Besides, unlike most other forms of colloquial Arabic, large resources of Egyptian Arabic can be found in written format in social Media. The difference between Egyptian Arabic dialect and MSA can be limited in certain phonological and morphological exchanges. For instance, Most of MSA nouns are preserved in EGY, but some other nouns have undergone some phonological changes such as:

- Monophthongization, e.g. /Sayf/ in MSA is turned into /Se:f/ in EGY.
- Final hamza deletion and final vowel shortening e.g. /sama:/ in MSA is turned into /sama/ in EGY.
- Atonic shortening e.g. /Sa:ru:x/in MSA is turned into /Saru:x/.
- Compensatory lengthening e.g. /ras/ in MSA is turned into /ra:s/.

Other critical difference between MSA and EGY is the case ending. Cases refer to what in English are called nominative, accusative, and genitive nouns. MSA distinguishes between the three cases by suffixing /u/ for nominative, /a/ for accusative, and /i/ for genitive [12]. However, case ending in EGY are deleted and they are understood by context, suffixes that are used to signify number is an additional concept that distinguishes EGY from MSA. In EGY, the masculine plural suffix /i:n/ is attached to masculine plural nouns, as well as, it is used for some feminine plural nouns beside the feminine suffix /a:t/ (Holes 2004, p. 166).

4 Machine Translation Paradigms

There are two main paradigms of MT: Rules-Based paradigms RBMT, and Statistical paradigms SMT. Hybrid machine translation is an approach to combine the achievements of both paradigms to reach to better results.

4.1 Rule Based MT

Rule based Machine translation uses the linguistic knowledge of the source language and the target language to accomplish the translation, rule based MT covers three main strategies: direct translation which translates word by word or linguistic patterns of the source language SL to others in the target language TL in a single step using bilingual dictionary, transfer system which based on contrastive knowledge to determine the differences between the two languages and it relies on creating rules to overcome these differences, transfer systems involve an analysis of the source text SL to an abstract structure the process that facilitates the transfer a corresponding abstract structure of the target text TL before generating it, Finally, interlingua which is divided into two phases: the analysis phase to encode the input text into interlingua and the generation phase to decode the interlingua into the output text [13, 14], interlingua systems require a transfer step as a part of the translation process [15, 16].

4.2 Statistical MT

It models the probability $P(F/E)$ of any source language F and target language E, the system chooses the translation that maximizes this probability. Initially it worked on the word level but later it is applied on larger chunks of the text [14, 17].

4.3 Hybrid MT

Since each paradigm has its strength and weakness, different approaches arose for combining these systems through hybridization such as:
Hybrid combination: to take one system and improve additional resource to enhance it, e.g. creating rules for a SMT, or vice versa.
Multi engine-Parallel combination: to translate using several independent systems.
Multi pass-Serial combination: to use the output of a system as an input of other system [14, 17].

Transfer systems are considered as a compromise between the ease use of the direct systems and the efficient uses of resources of interlingua systems [15]. The main advantage of the transfer systems is summarized in its ability to use a mediate language, in case of multilingual translation, into and out of which the translation is done [15].

5 (ALMoFseH) Arabic Dialects Machine Translation

The first English to Arabic Machine translation system was built in seventies by Weidner Communications Inc. it was composed of two main stages: the analysis system of the source language, and the generation system of the target language [18]. In this time and for decades after, there were no problem, since all the resourced were written in MSA [1]. After the appearance of the social media, texts which are written using a mixture of MSA and Arabic dialects were augmented, and understanding these texts became imperative. **(ALMoFseH)** project is an attempt to standardize the social media texts by identifying the different dialectal forms in these texts and transfer them into MSA. The project's goal is to develop a hybrid system, based on a multi pass serial combination, to translate the most used Arabic dialects in social media. Selecting these dialects is based on the annual report, that are released by Dubai school of governance and innovation program, to survey the Arabs usage of the social media. Building the system required serial processes, some are designed for the projects purpose and others relied on ready built applications. Hence, the system is planned to be composed of:

- Orthographic normalization to return the words that underwent changes owing to the Phonological Alternation rules into their origins.
- Dialect Identification by using an Automatic classifier to identify the dialectal forms in the text.
- Morphological analysis of the source text that written by the dialect using Egyptian Arabic Morphological analyzer.
- A Transfer system to select the lexical equivalents from dictionary lists. The system is to transfer from the Arabic dialects to MSA and from MSA to Arabic dialect, using machine learning classifier, rules and lexicon lists.
- A Morphological generator for the target dialect.

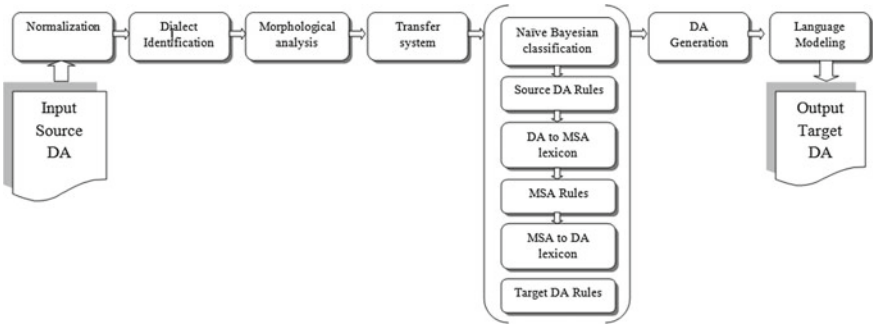

Fig. 1 The diagram shows the processing stages of developing Arabic dialects Machine translation using MSA as a pivot

- A Statistical language Model for the target dialect.

This research concerns with the procedures of creating A transfer system which is considered as the essential module in the Arabic Dialect Machine Translation system (Fig. 1).

6 Methodology

The work in this research is divided into two procedures: first, building a corpus based lexical database. Second, creating a transfer system for the selections from the Egyptian dialect to MSA.

6.1 Building a Lexical Database

A bi-dialectal lexicon is a crucial resource for building the application, since this lexicon is an essential resource to provide morpho-syntactic information such as POS, sub-categorization, tense, case, etc. [19]. Therefore, the objective of this stage is to develop a bi-dialectal lexical database extracted from social media texts such as Whatsapp and discussion forums. The process took different steps:

- Pre-processing of the corpus to clean the input from non-linguistics features
- Selection of most frequent words using a concordance
- Morphological annotating of the selected words
- Generating semantic information for the purpose of translation
- Developing a transfer module to be embedded in the proposed machine translation
- Evaluating the output

Table 1 Table 1 describes the components of the collected Corpus

Corpus componants	Percentage (%)
Non linguistic characters	11
Speech effects	6
Foreign words	18
Arabizi	23
Arabic words	42

A corpus of 250k words were gathered from the Whatsapp messages and the discussion forum. The data was composed of a mixture of Egyptian Arabic words and MSA word, Arabizi (Arabic words written by Latin letters), non-linguistics characters such as emoticons, and sound effects, as well as, the foreign words. The first step in preparing the data was to remove all those nonlinguistic characters. Sound effects were left in their original form since their spreading turned them into consistent standard linguistic forms that render certain meanings. Arabizi were transliterated by native speakers using their own writing style to maintain the realistic variations of the same word. Foreign words were the least among the other characters therefore they were removed. After filtering and manipulating the data the 178k words were listed according to their frequency using the word list in Antconc concordance. The most frequent 3k Egyptian Arabic words were sorted to develop our lexicon.

The developed lexicon is a stem-based lexicon which provides all acceptable stems for the individual word accompanied with their affixes. The lexicon is designed to accept other dialects to be added. Subsequently, the tool can be modified to be able to translate from dialect to other dialects using MSA. In this case, by extending the lexicon to establish a multidialectal lexicon, the input of any dialect D1 can be transferred to any other dialect Dn via MSA as an mediator. The lexicon is divided into 3 lists: the first list includes proclitics and prefixes, the second list includes enclitics and suffixes and the third list includes all possible stems for each word according to the adjacent affixes (Fig. 2).

The stem list is designed to provide all possible morphological and syntactic information for each word stem: category, tense, number, gender, voice, and tense. The list is divided into 3 main sub-lists:

```
( ['>a': 061], ,['prefcat':'1P']['num':'sing'], ['gen':'MASC']),
( ['ti': 062], ,['prefcat':'2P']['num':'sing'], ['gen':'MASC']),
 ( ['ti': 063], ,['prefcat':'2P']['num':'sing'], ['gen':'FEM']),
( ['ti': 064], ,['prefcat':'2P']['num':'PL'], ['gen':'MASC'])
( ['ti': 065], ,['prefcat':'2P']['num':'PL'], ['gen':'FEM']),
( ['yi': 066], ,['prefcat':'3P']['num':'sing'], ['gen':'MASC']),
 ( ['yi': 067], ,['prefcat':'3P']['num':'sing'], ['gen':'FEM']),
( ['yi': 068], ,['prefcat':'3P']['num':'PL'], ['gen':'MASC']),
 ( ['yi': 069], ,['prefcat':'3P']['num':'PL'], ['gen':'FEM']),
( ['yi': 070], ,['prefcat':'3P']['num':'sing'], ['gen':'FEM'])
```

Fig. 2 It shows a part of the prefix list

Non-inflected words list: it includes interrogative pronouns, relative pronouns, personal pronouns, demonstrative pronouns, prepositions, prepositional, adverbs, adverbials, pseudo verbs and non inflected verbs.
Inflected words list: this list contains EGY words that are descended from MSA but underwent orthographic deviations by altering a phoneme or more such as نَائِم and نَايِم, EGY words that have no origins in MSA such as تُورتَة whose equivalent in MSA is گَعگَة, OOV (Out Of Vocabulary) words, and borrowings that are taken from other language for specific purposes, and underwent the Arabic morphological inflection such as pluralization e.g. لَايگَات.
Broken plurals list: it stores the most frequent EGY Broken plurals with MSA broken plural equivalent e.g. فُلُوس and أَموَال, EGY sound masculine/feminine plurals with MSA broken plural equivalents e.g. سِتَّات and نِسَاء and EGY Broken plurals with MSA sound masculine/feminine plural equivalents e.g. أَزَايِز and زُجَاجَات.
3k Egyptian word types, and 3k MSA word types are the total number of the three lists that cover the following features:

- Orthographic variants: words that have several written forms such as بَرَدُو, بَرَضُو. All variants of each word type were inserted to the stem list.
- Words with no equivalent: some Egyptian words have no equivalent with the same meaning in MSA these word are called interjection and they are inserted to the sentence to express a reaction toward situation such as خَلَاص, بَقَا, طَبّ.
- Words with multi-word equivalent: some Egyptian words are translated to MSA using more than one word such as: مَعلِهش whose equivalent is لَا بَأس.
- New entries: lexical items that are now considered as a significant feature of the Egyptian dialect text such as the borrowings e.g. مِشيّر, and لَايكَات, and other new words such as, فَاكس. Each lexical item in the lexicon is enriched with morpho-syntactic and semantic information and a numerical code (Fig. 3).

6.2 The Transfer System

The process of transferring is the stage that follows the morphological analysis of the source dialect whose output are tokens with their POS. Thus, the target in this stage is to use output that resemble the Egyptian Arabic Morphological analyser CALIMA output [20] to reach to the closest MSA analyzed format of the Buckwalter morphological analysis for the equivalent MSA text [21]. The accomplishment of this process required three main procedures:

```
(['fas-aH':383], ['POS':'FV'],['cat':'FV']),
(['fas-aH':384], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':122]),
(['fas-aH':385], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':123]),
(['fas-aH':386], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':133]),
(['fas-aH':387], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':131]),
(['fas-aH':388], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':121]),
(['fas-aH':389], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':120]),
(['fas-aH':390], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':119]),
(['fas-aH':391], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_OBJ':128]),
(['fas-aH':392], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':122]),
(['fas-aH':393], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':123]),
(['fas-aH':394], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':131]),
(['fas-aH':395], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':121]),
(['fas-aH':396], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':120]),
(['fas-aH':397], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':119]),
(['fas-aH':398], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':128]),
(['fas-aH':399], ['POS':'FV'],['cat':'V'], ['tense':'past'], ['voice: 'act'], ['SUFF_SUB':111],['SUFF_OBJ':133]),
```

Fig. 3 It shows a part of the EGY stem list of the word /fas aH/ فسّح in the lexicon database

1. A machine learning classifier to select the mentioned token according to the context from the available analyses of the same word.
2. Rewrite rules to normalize the output tokens to suit the lexicon entries to facilitate looking up the tokens from the lists.
3. Looking up the normalized tokens from the lexicon with its all possible equivalents.

6.3 Naive Bayesian Classifier (NB)

Enhancing the system with learning methods to disambiguate the output of the morphological analyzer was urgently required. This stage guarantees the efficiency of the system and the avoidance of further undefined output. For this process, the supervised learning algorithm Naive Bayesian was chosen, since Naive Bayesian is the simplest representative of probabilistic learning methods' [22]. The description of the process of using NB is not the major subject of this research, therefore these lines are a brief overview of the usage of this method to reach to the desired target. According to NB the context in which the ambiguous words appear, is represented by vector of feature variables F = (f1, f2, …, fn), and the sense of these words are represented in a classification variable S = (s1, s2, …, sn). The disambiguation occurs through estimating the maximized sense according to the conditional probability P(w = si/F). Features in NB algorithms are terms as words, collocations, or words assigned by their position in the context [23]. The selected features in the system were:

F1 = a set of individual words,
F2 = a set of part of speech tags,
F3 = a set of words collocations,
F4 = a set of collocations of part of speech tags.

To choose the right sense of the ambiguous word in the given context, the conditional probability of the feature fi and the conditional probability of the sense si were computed using the Maximum Likelihood Estimation as follows:

$$P(s_i) = C(s_i)/N$$
$$P(f_i/w = s_i) = C(f_i, s_i)/C(s_i)$$

where $C(f_i, s_i)$ is the number of occurrence of the feature fi with certain sense si in the training corpus, $C(s_i)$ is the occurrence of this sense si in the corpus training, and N is the total number of the training corpus.

6.4 Rewrite Rules for Dialectal Normalization

After the tokenization and the morphological Analysis of the source dialect, some of the source words appeared differently from their saved forms in the lexicon. Definitely, in this case one of the main pre-processing stages was to normalize these words to assimilate their modified forms in the lexicon. The occurring modifications of these certain words were for the purpose of exceeding the changes between the source words and their equivalents in MSA. We gathered the most common words that fall under these conditions in sub-lists to facilitate writing rules for their normalization. These rules were designed to map the surface form of the word into the closest form to the words stored in the lexicon. These rules were categorized into 3 sorts:

Deletion Rules: for words with affixes that have no equivalent in MSA. Deletion rules were written to delete these affixes during the transfer.

```
Ex: ap → 0 || N[LIST04]_ 0

[HAj+NOUN+ap+NSUFF_SG :$iy>+NOUN+NULL]
```

This rule is designed to delete the suffix /ap/ that are joined to list of words in the source dialect and has no equivalent in MSA, such as /HAjap/ حَاجَة whose equivalent in MSA is /$iy/ شَيءٌ.

Alternation Rules: these rules are designed to alter certain morphemes with others that differ from their equivalents in the lexicon to match the referring meaning.

```
Ex: [N:1]→ [N:2]|| _[LIST05]

[EalaY+PREP: li+PREP]
```

The previous rule is written to alter the preposition /EalaY/ in the word /Eala/ عَلَشَان to the preposition li to avoid the wrong literal transferring.

Merging rules: these rules were written for individual cases when merging the analyzed morphemes is required to reach to the form existed in the lexicon.

```
Ex: [N:1]→[N:2]|| _ [LIST06]

[fiy+PREP : fiyh+PREP]
```

The word /fiyh/, فِيه is inserted to the lexicon without tokenization to match its equivalent in MSA /hunAka/, هُنَاك. Therefor the merging rule's role is to merge the two morphemes to match the lexicon entry. For instance, the Egyptian word /fiyh/ has two entries in the lexicon: the first entry renders the meaning (in it). And the second

Fig. 4 It shows one of the merging rules for the word (fiyh)

entry renders the meaning (there is). The rule is designed to cover the second entry. *Splitting Rules*: these rules were created for two purposes: first, to split the merged words in the source dialect, second, to split the words in MSA that stored in the lexicon with their affixes that have no equivalent in the source language.

```
Ex: [N:1] → [N:2]||N in (list06)

(bisml~Ah+NOUN : bi+PREP+Aism+NOUN+All~h+NOUN)
```

This rule is to split the two merged words and normalize them before the process of transferring. The most common merged words in the social media texts were gathered in a database list (list06) with their normalized form and their morphosyntactic information, as shown in the previous rule (Fig. 4).

Lexicon Look Up

Looking up in the systems lexicon follows certain restrictions to cover all the distinctions that distinguish the Egyptian dialect from MSA. As mentioned above, each morpheme in the source dialect lists have its own code number which matches another code number for the equivalent morpheme in the target dialect lists. Looking for the matched morphemes in both dialects were achieved in this stage by using the code numbers. Words in the source language that consists of one morpheme and whose equivalent consists of more than one morpheme has specific codes to facilitate the matching.

Fig. 5 Shows a part of the code of the case ending rule

6.5 Functional Model

One of the critical distinctions between the EGY and MSA is the case ending. Usually words in EGY dont include the case ending however, ending is understood by context. Therefore, output of the morphological analysis must be manipulated after transferring into MSA by adding the appropriate case ending. 38 rules were created to interpolate the main case endings that occurs persistently with verbs and nouns according to their tense, positions and definiteness. Part of compiling these rules using Python programming language is introduced here (Fig. 5).

7 Evaluation of the System

Evaluations of machine translations and their modules are needed to measure the performance of the systems by revealing haw far the output is accurate, predictable to the real human language. According to the general error metrics, the distance d(t, r) between the produced translation t and the predefined reference translation r is calculated and computed automatically.

An evaluation was conducted to measure the performance of the transfer system. The goal of the evaluation is: first, to measure the lexicon coverage of the Egyptian Arabic words through calculating the word error rate (WER). Second, to measure the accuracy of the output by comparing it to the output of the Buckwalter Morphological analyzer.

The first measurement was a measurement of the lexical database proficiency in the word level, each list (prefix, stem, suffix) in the built lexicon was measured separately. For that purpose, a blind test of 90 texts with total number of words 3000 words were collected from social media forums and SMSs, the data were collected

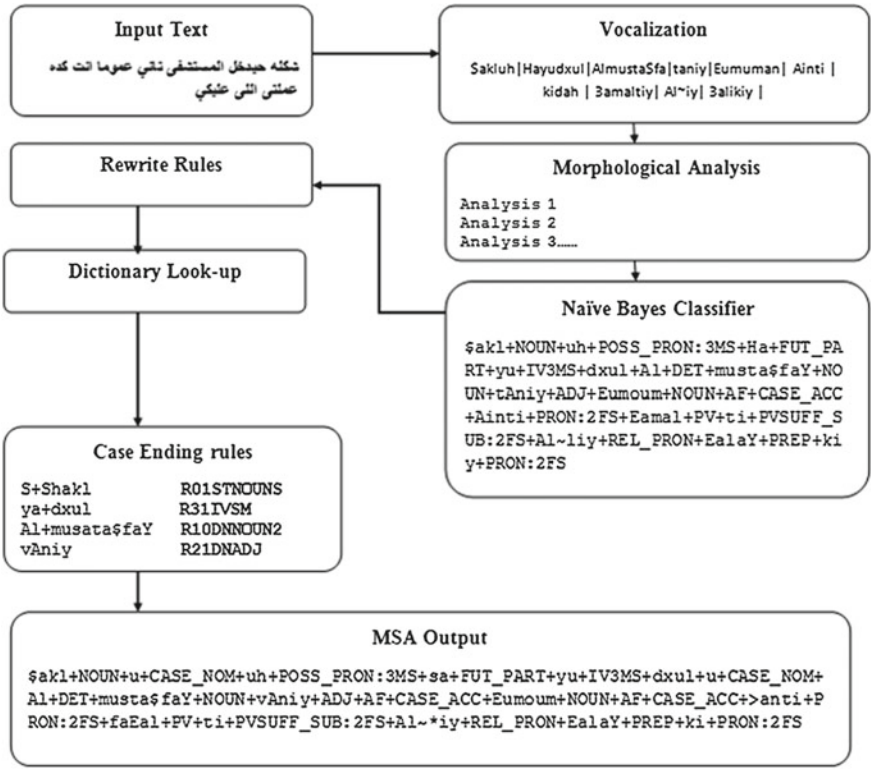

Fig. 6 Shows the process of transferring one of input sentences that are used for the evaluation

from different messages than those used in building the system's lexical database. Then it was analyzed by assistance of the morphological analyzer CALIMA to get the same format of its output. The output of the analysis was entered to the system manually. In this measurement, the Naive Bayes classification were excluded, since the aim of the measurement was to find out the lexicon ability to covers all the possible analyzed forms of the word. Thus, the error rate of each morpheme were calculated according to the following criteria: its existence in the lexicon, its existence in the lexicon with the same meaning according to the context, and the correctness of the equivalent morpheme.

The second measurement was designed to estimate the system's ability to matches the humans translation, and to measure the applicability of the system to provide an output that can act as a source input for a further transfer to other target dialect. Hence, the collected test data were manually translated into MSA by native Egyptian Arabic speakers. Then the MSA texts were morphologically analyzed using the open source of SAMA the last version of Buckwalter morphological analyzer. The output was sorted as database and compared with the systems output to measure the following values: Recall, Precision, F-score (Fig. 6).

8 Results

The results of the first evaluation shows a high coverage of the lexical features of Egyptian Arabic and the efficiency of the rules to facilitate finding the correct equivalent words. The accuracy of the system reached to 92.7%.

Due to the limited time and the number of the researchers who worked in this research, the size of the lexicon wasn't sufficient enough to cover a major number of the Egyptian words. That was the main reason behind the error rate in the evaluation.

Table 2, shows the word error rate (WER): first column presents the percentage of the morphemes that are transferred correctly due to the existence of the morpheme with its correct meaning according to the text, second column presents the morphemes that are not transferred correctly due to the existence of the word but with different meaning, and third column presents the morphemes that are not transferred from EGY to MSA and kept in its source form, due to the inexistence of the morpheme in the lexical database. The average of the correct rate is calculated for all the morphemes (Fig. 7).

The second measurement show that the output of the system could predict most of the analys feature and give an Approximate acceptable analyzed format for the target dialect MSA. Table 3, shows the recall, precision, and f score of the transfer system as a result of the second measurement.

Table 2 Shows the error rate of the lexical database

Feature	Correct (%)	Incorrect (%)	Untransfered (%)
Stem	90.8	2.1	7
Prefixes/Enclitics	95	2.4	2.5
Suffixes/Priclitics	93	7	–

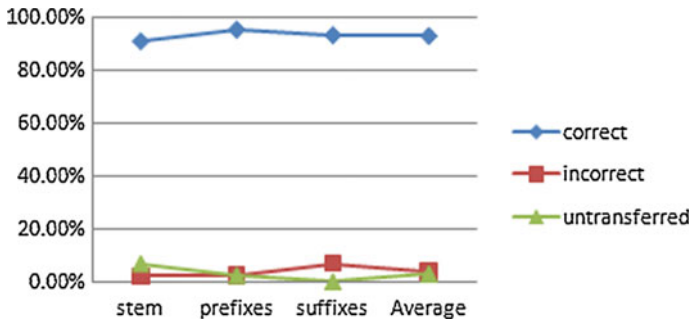

Fig. 7 Shows the error rate of the lexical database

Table 3 Shows the recall, precision, f-score

Feature	Recall (%)	Precision (%)	F-score (%)
Stem	88.1	82	84.8
Prefixes/Enclitics	98	96.8	98
Suffixes/Priclitics	91	89.1	90
Average	90	89	91

9 Conclusion

Arabic dialects Machine translation project (AlMoFseH) is a sophisticated project which demands various sequential processes to be drawn together. Each process should be manipulated separately to reach to satisfying results. This research exposed the paradigm that is used to accomplish one of these processes, and the problems that are confronted during building the transfer system and the procedures to handle them. The results of the primary work shows a high accurate performance of the transfer system. These results are encouraging to expand the work by increasing the database and the required rules before moving to the next stage of the project. For the future work, we would like to investigate the system's capability to transfer the MSA to other dialects by enlarging the lexicon to accept other dialects than the Egyptian dialect to reach to an approximate final phase of the project **(ALMoFseH)**. The project is planned to cover the Egyptian Arabic, Levantine Arabic, and Hijazi dialect.

References

1. Dasigi, P., Diab, M.T.: CODACT: Towards identifying orthographic variants in dialectal Arabic. In: IJCNLP, pp. 318–326 (2011)
2. Ibrahim, Z.: Beyond Lexical Variation in Modern Standard Arabic: Egypt. Cambridge Scholars Publishing, Lebanon and Morocco (2009)
3. Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O.F., Callison-Burch, C.: Machine translation of Arabic dialects. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 49–59. Association for Computational Linguistics (2012)
4. Bakr, H.A., Shaalan, K., Ziedan, I.: A hybrid approach for converting written Egyptian colloquial dialect into diacritized Arabic. In: The 6th International Conference on Informatics and Systems, infos2008. Cairo University (2008)
5. Monem, A.A., Shaalan, K., Rafea, A., Baraka, H.: Generating Arabic text in multilingual speech-to-speech machine translation framework. Mach. Transl. **22**(4), 205–258 (2008)
6. Sawaf, H.: 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)
7. Salloum, W., Habash, N.: Elissa: a dialectal to standard Arabic machine translation system. In: COLING (demos), pp. 385–392 (2012)

8. Holes, C.: Modern Arabic: Structures, Functions, and Varieties. Georgetown University Press (2004)
9. El-Hassan, S.A.: Educated spoken Arabic in Egypt and the Levant: a critical review of diglossia and related concepts. Archivum Linguisticum Leeds **8**(2), 112–132 (1977)
10. Bani-Khaled, T.A.A.: Standard Arabic and Diglossia: a problem for language education in the Arab World (2014)
11. Hassig, H.: Deriving Cairene Arabic from modern standard Arabic: a framework for using modern standard Arabic text to synthesize Cairene Arabic speech from phonetic transcription (2011)
12. Gadalla, H.A.: Comparative Morphology of Standard and Egyptian Arabic, vol. 5. Lincom Europa Munich (2000)
13. Okpor, M.: Machine translation approaches: issues and challenges. Int. J. Comput. Sci. Issues (IJCSI) **11**(5), 159 (2014)
14. Shilon, R.: Transfer-based machine translation between morphologically-rich and resource-poor languages: the case of Hebrew and Arabic. Ph.D. thesis, Citeseer (2011)
15. Trujillo, A.: Translation Engines: Techniques for Machine Translation. Springer Science & Business Media (2012)
16. Hutchins, W.J., Somers, H.L.: An Introduction to Machine Translation, vol. 362. Academic Press London (1992)
17. Artetxe Zurutuza, M.: Distributional semantics and machine learning for statistical machine translation (2016)
18. Farghaly, A.: Arabic machine translation: a developmental perspective. Int. J. Inf. Commun. Technol. **3**, 3–10 (2010)
19. Tze, L.L.: Multilingual lexicons for machine translation. In: Proceedings of the 11th International Conference on Information Integration and Web-based Applications and Services, pp. 734–738. ACM (2009)
20. Habash, N., Eskander, R., Hawwari, A.: A morphological analyzer for Egyptian Arabic. In: Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, pp. 1–9. Association for Computational Linguistics (2012)
21. Habash, N., Diab, M.T., Rambow, O.: Conventional orthography for dialectal Arabic. In: LREC, pp. 711–718 (2012)
22. Escudero, G., Màrquez, L., Rigau, G.: Machine Learning Techniques for Word Sense Disambiguation (2003)
23. Le, C.A., Shimazu, A.: High WSD accuracy using Naive Bayesian classifier with rich features. Proc. PACLIC **18**, 105–113 (2004)