Toward Building a Comprehensive Phrase-based English-Arabic Statistical Machine Translation System

Sara Ebrahim[*1], Samhaa R. El-Beltagy[**2], Doaa Higazy[*3], Mostafa G. M. Mostafa[***4]

[*]*Scientific Computing Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt*
[1]sara.elkafrawy@gmail.com
[3]doaa.hegazy@cis.asu.edu.eg

[**]*Centre for Informatics Science, Nile University, 6 October, Egypt*
[2]samhaa@computer.org

[***] *Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt*
[4]mgmostafa@cis.asu.edu.eg

Abstract: ***This paper explores a phrase-based statistical machine translation (PBSMT) pipeline for English-Arabic (En-Ar) language pair. The work surveys the most recent experiments conducted to enhance Arabic machine translation in the En-Ar direction. It also focuses on free datasets and linguistically motivated ideas that enhance phrase-based En-Ar statistical machine translation (SMT) as it is as aims to use those only in order to build a large scale En-Ar SMT system. In addition, the paper highlights Arabic linguistic challenges in Machine Translation (MT) in general. This paper can be considered a guide for building an En-Ar PBSMT system. Furthermore, the presented pipeline can be generalized to any language pairs.***

Key words: ***Machine Translation, Arabic Natural Language Processing, Phrase-based, Statistical Machine Translation.***

1 INTRODUCTION

Developing an automatic Machine Translation (MT) system over the history poses many challenges to researchers. It has been worked on since the World War I when there was lack of human translators and the need of instance translation was highly needed. Machine Translation has been tackled with various techniques such as:

- Direct approach: It is the first type of MT to appear. It is called word-for-word translation. It is more likely to be a bilingual dictionary; each word in the source language is looked after in the dictionary to come up with the corresponding word in the target language. The process was divided into three steps:
  1) Pre-processing the source text: analyze source text morphologically and extract the lemma forms.
  2) Dictionary look up: find the translation of a single source word in a target language dictionary.
  3) Final output: generate the whole sentence after looking up each word separately.

  There is an obvious drawback of this approach which is neglecting the sentence connections in the translation process. This suggests a more inter-processes in the three aforementioned steps.

- Transfer approach: It aims at transferring the source text into the target text through a middle-ware. The middle-ware is the syntax analysis of both languages. The steps can be formulated as following:
  1) Analyze: Analyze the source text and parse it in order to get its parse tree.
  2) Transfer: Transfer the source text parse tree into a new parse tree for the target language.
  3) Generate: Generate the target text from the new parse tree.

- Interlingua approach: It is trying to find a universal language that any language can be translated into. This universal language aims to be independent and an intermediate between source and target texts. However, this approach' idea is to represent the semantic analysis of the source text in an abstract logical form.

- Statistical approach: It does not require prior linguistic knowledge; and this is the most important advantage of this approach. Statistical Machine Translation (SMT) is a promising direction in MT field; with the available huge amount of parallel data (translated documents) statistical models can be trained efficiently to translate between any two language pair. In this paper we are going to focus on SMT among other methods in MT field. SMT is an MT paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text named as parallel corpora. Despite the fact that SMT is widely used more than other paradigms it has shortcomings, some of them are:

- Corpus creation can be costly for users with limited resources.
- The results are unexpected. Superficial fluency can be deceiving.
- The benefits are overemphasized for European languages.

2 STATISTICAL MACHINE TRANSLATION SYSTEM PIPELINE

To the best of our knowledge, there are currently no surveys that explore state of the art linguistic ideas, tools and available datasets for building an En-Ar SMT system. While Ebrahim et al. in Ref. [15] surveyed different modifications that contribute to the enhancement of En-Ar SMTs, they did not review available free tools and datasets in order to conduct real experiments. The aim of this work is to present such a survey.

SMT systems usually fall under one of two categories: phrase-based models or tree-based models (hierarchical phrase-based and syntax-based). SMT for either category, has a pipeline which is identical for any pair of languages; in other words, it is a linguistically independent pipeline. Enhancing specific linguistics features has been shown to boost the automatic evaluation scores. This survey lists the Arabic, the target language of this survey, linguistically motivated ideas for En-Ar SMT system. This paper will focus on phrase-based SMT systems and the linguistics enhancements proposed for Arabic.

As illustrated in figure 1, an SMT system starts with a parallel text for the language pair. The parallel text (also known as parallel corpus) should be aligned on a sentence level; each line in one of the two files representing the language pair has its translation in the corresponding line number in the other file. In addition to the parallel corpus the pipeline needs a monolingual corpus for the target language to train the language model (in our case it will be for the Arabic language).
The idea of phrase-based SMT is summarized in three steps:

1. Create a lexicon of parallel phrases.
2. Calculate (estimate) the score or the probability of each possible translation for each phrase.
3. For a new sentence, search for the translation that obtains the highest score. This process is called decoding.

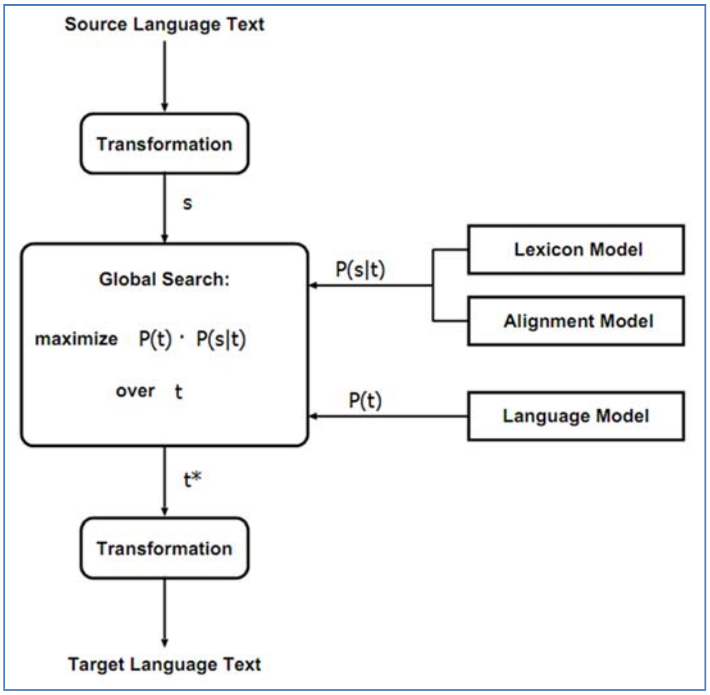


Figure 1 SMT System Architecture

The first step in the pipeline is the alignment process. The parallel corpus is aligned using historically known IBM alignment models Ref. [39]. The alignments are extracted from an intersection of bidirectional alignments (En-Ar and Ar-En), in addition to some union alignments of the two processes. At this point, it is easy to extract a maximum likelihood lexical translation table. The lexical translation table has each word/phrase, a possible translation for that entry, and the probability of that translation. The result is a table that has all phrases and the corresponding translations and this table is called the phrase table. The alignment process shown in figure1 can be named phrase table extraction process. A detailed, easy to understand step-by-step explanation of the alignment idea, can be found in a workbook published on the website of the Information Science Institute (ISI) at South California University[1].

A language model (LM) is also trained using the monolingual corpus of the target language (the Arabic language in this survey). The LM indicates the extent to which the resultant target sentence is actually a valid Arabic language sentence, while the translation model (TM) (i.e.) is trained with the parallel En-Ar corpus. After training the language model from the monolingual corpus and TM from the parallel corpus, new sentences are ready to be translated by the system. Both models, LM and TM, inherit from the noisy channel mode. LM can be a trigram model, a factored model, or other types. The noisy-channel model uses Bayes rule and illustrated in the context of SMT in equation (3).

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

Hence,

$$\operatorname*{argmax}_{t \in A} p(t|s) = \operatorname*{argmax}_{t \in A} \frac{p(t)p(s|t)}{\sum_t p(t)p(s|t)} \tag{2}$$

$$p(t|s) = \operatorname*{argmax}_{t \in A} p(s|t)p(t) \tag{3}$$

Then, the decoder searches for the best translation for each sentence, which is a set of phrases. In other words the decoder searches for the translation that has the highest probability. The decoder should (in an ideal case) extract every possible translation; it should take each word and search for its best match, then take the second word and find the possible routes to achieve the highest translation score. This ideal case is impossible with current computing resources as the number of available phrases increases exponentially with respect to matched entries in the phrase table. Instead, the decoder does a heuristic search by discarding less promising hypotheses which allows the search process to be feasible.
The decoding task or the search task aims to find the target sentence (Arabic) that has the highest probability after calculating the product of the LM score and the TM score. Mathematically, this task of searching for a new Arabic sentence, can be formulated as:

$$t^* = \operatorname*{argmax}_{t \in A} \; p(s|t) \; \times p(t) \tag{4}$$

All sentences in Arabic is the set denoted as in equation 4. Hence, a potential translation score is the product of two scores: the LM score, gives a prior distribution for which sentences are likely to be valid Arabic, and the TM score, which indicates how likely an English sentence can be as a translation of the target sentence.

The rest of this paper is organized as follows: an exploration of Arabic linguistics challenges is presented in the next section, section three lists available corpora(monolingual and parallel) for training an SMT, section four has the language independent tools and frameworks in order to implement the SMT pipeline. Finally last section, the core of this paper, has the linguistic tools and techniques that proved to enhance En-Ar SMT evaluation scores.

3 ARABIC CHALLENGES IN MT

Despite the fact that SMT systems have shown reasonable results in close language pairs, the same results were not achieved for distant language pairs. The fact that English to Arabic is a distant pair has meant that SMT systems targeting that pair, have achieved unsatisfactory results. Moreover, Arabic has its own challenges within the Natural Language Processing (NLP) field. In the following sub-sections, the most important of those challenges, are presented.

A. Orthographic Challenges

Arabic has a complex orthography when it comes to computational linguistics; for example, in the Optical Character Recognition(OCR) field, scientists face the problem of connected Arabic letters which is very different from English where each letter is separated from the others by a space. Moreover, a letter in Arabic can have three different forms depending on whether it appears at the beginning, in the middle or at the end of a word.

Having complex orthography, it is difficult for majority of writers to produce a correct Arabic form. For example, it is confusing for many to differentiate between (“علَى” which means the preposition *on*) and (*"*علي*"* which means a proper name Ali). This issue increases the sparsity (i.e. having many forms of the same word) and ambiguity (i.e. a word has

[1] www.isi.edu/natural-language/mt/wkbk.rtf

multiple forms) in SMT language model and translation model training. In particular, various forms of Hamzated Alif “ آ، أ، إ“exist in almost all Arabic scripts with no Hamza "ا". Two forms of Ya are also used incorrectly at the end of a word: the dotless Ya (ى) or the Alif-Maqsura and the dotted Ya (ي) Ref. [17].[2]

B. Morphological Challenges

Arabic is a morphologically complex language. Compared with the English language, Arabic is richer than English morphologically. Arabic words are inflected for number and gender, and can be attached to different clitics such as:

- conjunction(*w* + means: *'and'*).
- possessive pronouns and object pronouns(e.g. + *ny means: 'me/my'*, *+hum means: 'their/them'*).
- the definite article(*Al*+ means: *'the'*).
- preposition(e.g. *b*+ means: *'by/with'*.
- *k*+ means: *'as'*).
- *l*+ means: *'for'*.

As an example, the nominal phrase *wbsyyArAtnA* وبسيارتَنا, and the verbal phrase *wsnkAtbhum* وسنُكاتبهم are cliticized as following:

1. w+ s+ n+ kAtb+ hum
and+ will+ we+ write+ to them
2. w+ b+ syyAr+ At+ nA
and+ with+ car+ PL+ our

The richness in Arabic morphology leads to having many surface forms in the parallel corpus, when compared to the English side and the sparsity problem appears. El Kholy and Habash in Ref. [18] said that While the number of (morphologically untokenized) Arabic words in a parallel corpus is 20% less than the number of corresponding English words, the number of unique Arabic word types is over twice the number of unique English word types over the same corpus size. [3]

C. Syntactical Challenges

Arabic syntax is more complex than English syntax. Among many syntactical issues in the Arabic language in the NLP field, three issues appear in the MT field: the adjectives, the verb-subject order and Idafa construct (equivalent to the English possessive, of-relationship, and compound nouns). Illustrative examples in this section are from Ebrahim et al. in Ref. [15].

1) Arabic Adjectives

The structure of the noun phrase in Arabic is different than English; the Arabic adjective that modifies a noun agrees with the noun in definiteness, thus a definite article is added to it if the noun is definite and vice versa:

1. Alyd Alkbyra

   the hand the big

   En: The big hand.

   Ar: اليد الكبيرة

2. yd kbyra

   hand big

   En: A big hand.

   Ar: يد كبيرة

2) Verb-subject Order

Closed language pairs seem to have similar structure order. Since Arabic is distant from English, it has a different order: Verb-Subject-Object order(VSO), while English has the Subject-Verb-Object (SVO) order. SVO order exists in Arabic but with lower frequency than VSO order and is not preferable. Examples in (1) and (2) illustrate

[2] All Arabic transliteration are provided in the Habash-Soudi-Buckwalter transliteration scheme Ref. [27].

[3] Illustrative Arabic Examples and the problem definition mentioned in this section are adapted from Badr et al. Ref. [6].

different ordering in Arabic. Example (3) illustrates the gender agreement in VSO order between the verb and the subject while example (4) illustrates the agreement in verb-subject gender and number in SVO order.

1. ktb Alwld Aldrs

   wrote the boy the lesson

   En: The boy wrote the lesson.

   Ar: كتب الولد الدرسَ

2. Alwld ktb Aldrs

   the boy wrote the lesson

   En: The boy wrote the lesson.

   Ar: الولد كتب الدرس

3. ktb AlA'wlAd Aldrws

   wrote the boys the lessons

   En: The boys wrote the lessons.

   Ar: كتب الأولاد الدروس

4. AlA'wlAd ktbw Aldrws

   the boys wrote the lessons

   En: The boys wrote the lessons.

   Ar: الأولاد كتبوا الدروس

3) *Idafa Construct*

The Idafa construct in Arabic is the equivalent version of the English possessive, *of*-relationship, and compound nouns. The translation of the three structures are the Idafa construct in Arabic, which contains one or more indefinite nouns then a definite noun. The English phrases (*the student books*, *the student's books* and *the books of the student*) for example are translated into one Arabic phrase which is (ktb AlTAlb – كُتُب الطالب).

4 AVAILABLE CORPORA

Third world universities lack funding in many fields including NLP and SMT. Datasetavailability was crucial for us to be able to carry out practical experiments. While many such dataset are avialable for purchase, lack of funding has forced us to search for freely available to use datasets for En-Ar SMT bearing in mind that an SMT system needs both a monolingual corpus and a parallel corpus to train both the language model and the translation model. In the following section, we will introduce the free datasets for En-Ar SMT that we were able to locate.

A. *The United Nations Parallel Corpus*

In 2009 Rafalovitch et al. in Ref. [46] published a parallel corpus for six language extracted from United Nations (UN) documents, and this corpus can be downloaded in different formats[4]. This corpus was extracted from the translation memories of the UN by individual researchers and it was not officially published by the UN.

Then, in 2010 Eisele et al. in Ref. [16] described the extraction process of the UN documents. They discussed methods used for crawling and formatting documents as well as for sentence alignment. Moreover, they provided a test set that can help in the evaluation of an SMT system. The paper is available for reading, but at the time of writing this paper, the corpus download page is encountering an error[5].

In 2016 an official UN parallel corpus was released Ref. [53]. The new corpus was published in the six official UN languages and was sentence aligned. Moreover, the authors provided the parallel corpus with development and test sets that can be used in any SMT system. The En-Ar parallel corpus contains 111,241 files with 18,539,207 lines. Most SMT systems targeting English-Arabic transaltion, employed a a smaller number of lines for training. The size of the new UN

[4] http://www.uncorpora.org/

[5] http://www.euromatrixplus.eu/downloads

corpus is promising for building future En-Ar SMT systems because the quality of SMT systems often relies heavily on the size of the training corpora in boththe language model and the translation model. The new UN corpus is also known as UNv1.0 corpus and can be downloaded in different formats[6].

B. The Linguistic Data Consortium Corpora

The Linguistics Data Consortium(LDC) is an institution that publishes a wide set of language resources periodically[7]. LDC is rich in parallel and monolingual corpora. Eventhough these corpora are not free LDC provides an application for a data-scholarship that once granted, allows its user to access the datasets freely. The data-scholarship program is offered twice a year, the first round takes place mid September while the second round is in mid January[8].

With respect to the task of EN-AR SMT, LDC has released the fifth edition of the Arabic Gigaword corpus, which is the most widely used monolingual corpus to train the language model in any Arabic SMT research paper. The Arabic Gigaword corpus, consists of Arabic news articles collected from nine online news sources (such as: Asharq Al-Awsat, Agence France Presse ... and Al Hayat)[9]. The fourth edition of the same corpus has 848469 separated tokens in 2716995 documents[10].

LDC has also released a number of Ar-En parallel corpora examples of which are those with the following catalog numbers: LDC2014T03, LDC2014T08, LDC2014T19, LDC2014T22, LDC2014T05, LDC2014T10 and LDC2014T14, LDC2013T14, and LDC2013T10[11]. In 2007, LDC released an automatically extracted parallel dataset LDC2007T08[12]. If this corpus results in improved evaluation scores in any En-Ar or Ar-En SMT systems, this will be a huge progress towards MT; because it means that building an En-Ar SMT system will not need a human-translators in order to have a ready parallel corpus. However, this has yet to be investigated.

In order to evaluate an SMT system, a set of source text and human translation references are required. LDC published a number of evaluation sets such as: LDC2014T02, LDC2013T07, LDC2013T03, LDC2010T10, LDC2010T11, LDC2010T12, LDC2010T14, LDC2010T17, LDC2010T21, LDC2010T23 and LDC2010T01. These datasets are all availed by NIST OpenMT[13]. According to the official website, NIST OpenMT is an evaluation series that supports research in, and helps advance the state of the art of, machine translation (MT). Most of the NIST evaluation sets target evaluating Ar-En systems To overcome this issue in evaluating an En-Ar SMT system, some researchers duplicate the Arabic translation four times in order to have more references and get an efficient automatic evaluation score.

C. Abu El-Khair Corpus

Ibrahim Abu El-Khair published in Ref. [19] Abu El-Khair Corpus, which is a Modern Standard Arabic Corpus. Abu El-Khair corpus reported to have more than five million news paper articles and a billion and a half words total.

5 Language Independent Tools

The phrase-based statistical machine translation (PBSMT) system trains using aligners and language model creators. Then, a decoder uses the trained models in order to translate new sentences. Since SMT tends to be language independent, most of the papers experiments' results are reported with automatically evaluated scores. In this section, we will explore the language independent tools for building a PBSMT system and the different automatic evaluation scores.

A. Translation Model Generators

GIZA++ is the most popular free parallel corpus aligner Ref. [39][14]. MGIZA is the multi-threaded version of GIZA++ and is kept up to date to work with most compilers[15]. In 2016 Cadigan et al. Ref. [8] released a distributed computed version of GIZA++ implemented over Spark by Apache; its speed is reported to be up to 5.6x that of GIZA++ and 2.6x that of the multi-threaded version of MGIZA.

[6] https://conferences.unite.un.org/UNCorpus
[7] https://www.ldc.upenn.edu/
[8] https://www.ldc.upenn.edu/language-resources/data/data-scholarships
[9] https://catalog.ldc.upenn.edu/LDC2011T11
[10] https://catalog.ldc.upenn.edu/LDC2009T30
[11] https://catalog.ldc.upenn.edu/byyear
[12] https://catalog.ldc.upenn.edu/LDC2014T10
[13] http://www.itl.nist.gov/iad/mig/tests/mt/
[14] https://github.com/moses-smt/giza-pp
[15] https://github.com/moses-smt/mgiza

Berkeley has an active research group in the area of NLP with many of its projects and tools published on its official website[16]. One of those tools is the **BerkeleyAligner** which is a word alignment software package that implements novel algorithms for unsupervised word alignment[17].

Anymalign is another aligner that is described by Lardilleux et al. in Ref. [36]. One of its reported main advantages over similar tools is that it can align any number of languages simultaneously[18]. Finally, there is **Chaski** which is a distributed PBMT training tool based on Hadoop[19].

B. Language Model Generators

Heafield Ref. [30] described the implementation of **KenLM**, a language model which is reported to have smaller and faster language model queries[20]. In benchmarking experiments done with Europarl parallel corpus, KenLM was reported to outperform the speed of BerkleyLM Ref. [45] by 4.49x[21]. KenLM was later integrated with Moses Ref. [34] (Moses is SMT decoder, and a comprehensive SMT system that will be discussed later in this section). KenLM is LGPL licensed (i.e. available for commercial use).

SRILM Ref. [49] is a toolkit for applying statistical LMs. SRILM is used in SMT and other NLP subfields. It has been under development in the SRI Speech Technology and Research Laboratory since 1995. SRILM is free to use in projects that do not receive external funding other than government research grants and contracts[22].

IRSTLM, like KenLM, is an LGPL licensed toolkit for generating statistical LMs, and is available for commercial use[23].

BerkleyLM is a library for storing large n-gram LMs efficiently in memory. The BerkleyLM is described in Ref. [45], and is reported to be fasterthan SRILM and nearly as fast as KenLM despite the reported results in the KenLM benchmark experiments[24].

RandLM Ref. [50] is an LM that uses randomized data structures, which is different from both SRILM and IRSTLM. The tool is recommended for use by Moses in its official web page when a user wants to build the largest LMs possible (e.g. a 5-gram on one hundred billion words). The result can be ten times smaller LMs than other LM toolkits. Talbot and Osborne described the technical details of RandLM in Ref. [50]. RandLM can be downloaded from Sourceforge[25].

In 2013 Vaswani et al. Ref. [52] published a Neural Probabilistic Language Model toolkit (NPLM)[26]. Then, in 2014, Paul et al. described Ref. [44] a neural network LM framework for machine translation which they called OxLM(Oxford LM). The framework can be downloaded from Sourceforge[27].

C. Decoders

Moses is an Open Source Toolkit for Statistical Machine Translation Ref. [34]. An SMT framework is more a precise description for Moses; because most of the mentioned LMs and translation model training toolkits are integrated within Moses. Moreover, Moses has been technically supported since 2005. It was built by researchers for research, has an active support mailing list and its code is shared on github. Different institutions contributed to Mosesâ€™s development including: University of Edinburgh (UK), Fondazione Bruno Kessler (Italy), Charles University (Czech Republic), DFKI (Germany), RWTH Aachen (Germany) and others. Its website is a wide gate towards understanding, implementing and enhancing SMT systems[28]. There is a documentation on the website that illustrates the installation steps and usage for all integrated tools. Moses is also available for commercial use because it is LGPL licensed. In addition, Moses has an experimental management system described in Ref. [35], which automates the whole SMT pipeline (illustrated in figure 2), and the workflow is automatically generated (e.g. in figure 3).

[16] http://nlp.cs.berkeley.edu/software.shtml
[17] https://code.google.com/archive/p/berkeleyaligner/
[18] https://anymalign.limsi.fr/
[19] https://sourceforge.net/projects/chaski/
[20] http://kheafield.com/code/kenlm/
[21] http://kheafield.com/code/kenlm/benchmark/
[22] http://www.speech.sri.com/projects/srilm/
[23] https://sourceforge.net/projects/irstlm/
[24] https://code.google.com/archive/p/berkeleylm/
[25] https://sourceforge.net/projects/randlm/
[26] http://nlg.isi.edu/software/nplm/
[27] https://github.com/pauldb89/OxLM/blob/master/README.md
[28] http://www.statmt.org/moses/

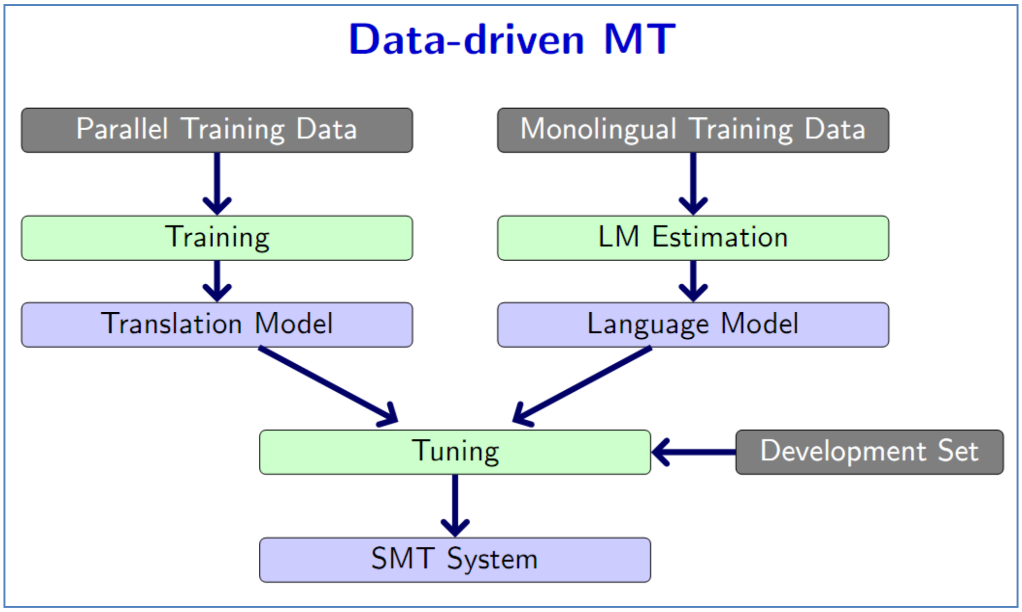


Figure 2: Moses Pipeline in a Phrase-Based SMT System. Ref. [51]

A transliteration model recently integrated with Moses, and was described and implemented by Durrani et al. Ref. [13]. A transliteration tool is important in Arabic SMT; because Arabic is not written in roman characters. This causes the increase of out of vocabulary words (OOVs), and transliteration can help reduce OOVs by transliterating Named Entities (NEs).

Thot is a toolkit for PBSMT. Ortiz et al. published Ref. [41] the new interactive toolkit. The new toolkit comes with number of improvements such as: Integration with a set of pre/post-processing tools, increased portability(compiled in many different platforms), improved checking technique for runtime errors, early detection of bugs using built-in checks, and translation can be executed in parallel either through multi-threading or distributed-computing paradigms. It has a detailed and reviewed manual[29].

Developed by Stanford, **Phrasal** is another phrase-based SMT decoder written in Java. The details of the Phrasal decoder are described in Ref. [2].

Docent is a document-level phrase based SMT Ref. [3]. It is worth noting that the Docent team acknowledges the work in Moses and KenLM for producing the Docent system[30].

Dyer et al. Ref. [14] described **cdec**, which is a decoder, aligner and a model optimizer for SMT based on context-free formalisms.

GREAT Ref. [23] Ref. [24] is a decoder based on stochastic finite-state transducers, which includes a training toolkit. Gonzalez et al. described the latest enhancements to the GREAT decoder in Ref. [25]. The research lab that published GREAT has a description of other MT tools on its website[31]. Interactive GREAT (iGREAT) is available for download[32].

Marie is an ngram-based SMT decoder developed in 2006 by Josep M. Crego as part of his PhD thesis Ref. [37]. The decoder details were published in the Computational Linguistics Journal with the title "N-gram-based machine translation" and the tools are available for research purposes[33].

Phramer is an open-source statistical phrase-based machine translation decoder that was released in 2006 Ref. [40] and is available for downloaded[34].

[29] http://daormar.github.io/thot/
[30] https://github.com/chardmeier/docent/wiki
[31] https://www.prhlt.upv.es/page/software
[32] https://sourceforge.net/projects/igreat/
[33] http://www.talp.upc.edu/index.php/technology/tools/machine-translation-tools/75-marie
[34] https://sourceforge.net/projects/phramer/

Pharaoh is another machine translation decoder for phrase-based systems released to the research community in 2004 Ref. [32]. It aimed to aid research in SMT. It was developed by Philipp Koehn as part of his PhD thesis at the University of Southern California and the Information Sciences Institute. It is worth noting that Pharaoh is the first phrase based SMT decoder and Philipp Koehn is the founder and a main contributor to the Moses system community that was mentioned earlier in this section.

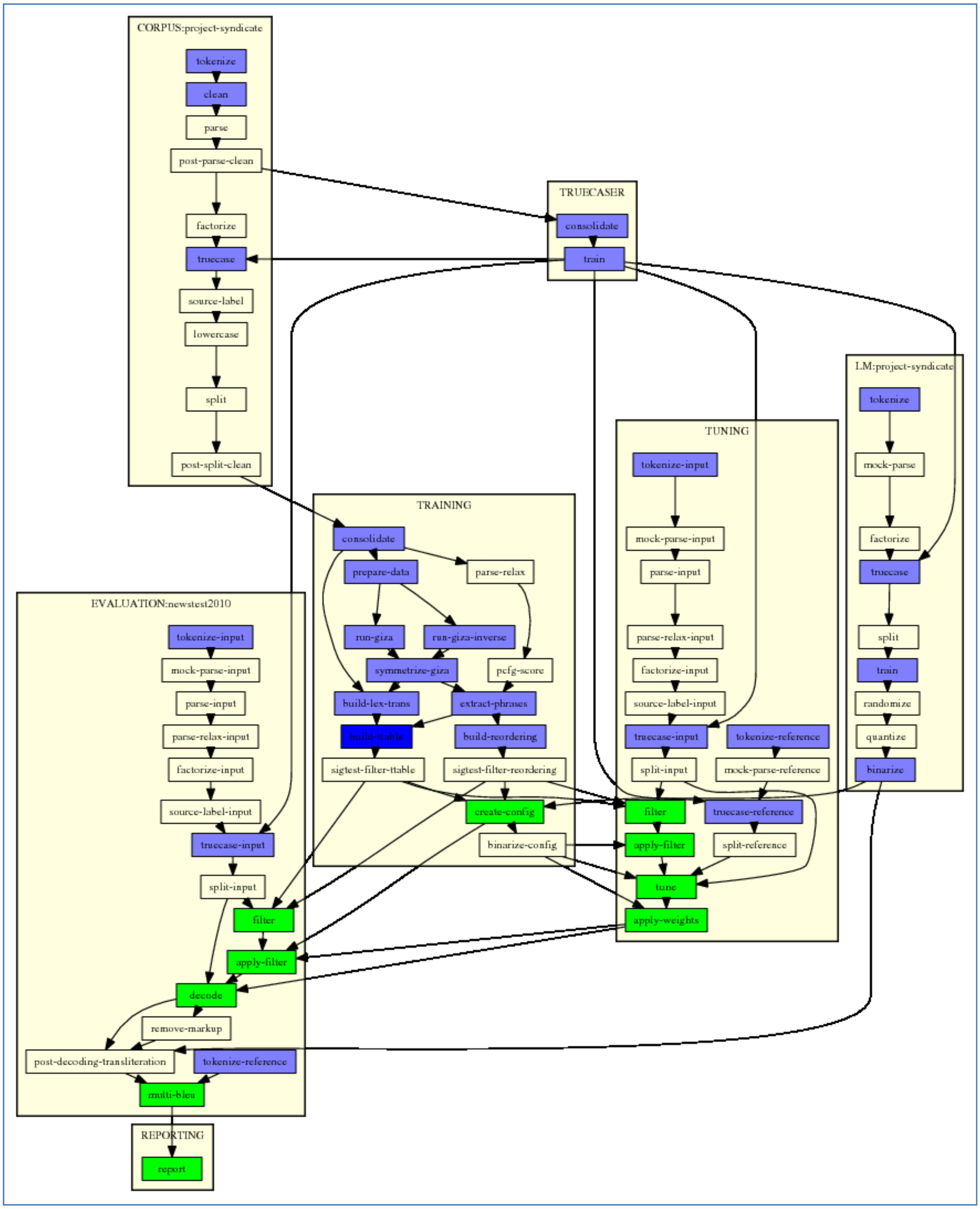


Figure 3 EMS workflow generated by Moses

D. Automatic Evaluation Metrics

Despite the fact that automatic evaluation for SMT is still a controversial topic, automatic evaluation scores are frequently used in reporting experiment results. BLEU (BiLingual Evaluation Understudy) was the first introduced automatic evaluation score Ref. [42]. It is a quick and language-independent score. It relies on a numerical metric for translation closeness; as it is believed that, the closer an MT is to a human translation, the better it is. In addition, it relies on good quality human translations (called references) of the test corpus.

There are other automatic evaluation metrics such as: Translation Error Rate (TER) which measures the number of edits required to change a system output into one of the references[35], METEOR[36] and RIBES[37] metrics.

E. Manual Evaluation Metrics

MT human evaluation is an topic as it is often considered the most reliable way for evaluating a MT system. . However manual evaluation is often very costly. This is the reason that motivated Chatzitheodorou Ref. [10] to release COSTA MT, which is an evaluation tool and an open-source Java software that can be used to facilitate the manual evaluation of MT output. As is reported, COSTA is simple to use and is designed to allow developers and users of MT systems analyze their engines within a friendly environment. It ranks the quality of MT system output segment-by-segment for a particular language pair[38].

Appraise Ref. [21] is another open-source tool for manually evaluating MT output. the author of Appraise has described it as a tool that allows the collection of human evaluations on translation output, implementing annotation tasks, and manual post-editing Ref. [21]. Appraise has also been used in the ACL WMT evaluation campaign.

6 TOOLS FOR LINGUISTICS ENHANCEMENTS

This paper focuses on SMT for the English-Arabic direction, even though SMT tends to be language independent. The reason for this is that we are interested in vitalizing the Arabic language as it was approved by many studies that learning with the native language is more effective and enhances creativity.

In Ref. [15], Ebrahim et al. state that machine translation in the Ar-En direction has more funding institutions than in the En-Ar direction. This statement is reported by Farghaly and Shaalan Ref. [20] who have explained that the need to understand what is said and written in Arabic has risen significantly after the event of September 11th, 2001. The applies to communication in airports, text messages and via telephone calls, and there was a lack of human translators. This fact was also stated Koehn Ref. [33] who has said that Due to the involvement of US funding agencies, most research groups focus on the translation from Arabic into English and Chinese into English. Next to text-to-text translation, there is increasing interest in speech-to-text translation.

It is logical that any enhancements that improve the Ar-En direction should have an impact on En-Ar. This proved to be true by many studies aimed at improving Ar-En by processing the Arabic language on different levels (i.e. orthographically, morphologically, and syntactically). There are different processing methods for the Arabic corpus: Morphological tokenization/detokenization, orthographic normalization/denormalization, syntactic reordering and Part of Speech Tagging(POS). On the other side of the corpus, processing the English language proved to help in SMT efficiency such as: POS, down-casing, cleaning (e.g. adding spaces around punctuation) and Named Entity Recognition(NER). In this section, we will explore most recent studies targeted processing Arabic SMT and tools to perform them.

A. Orthographically Processing Techniques

1) Orthographic Normalization

Orthographic Normalization is an Arabic text pre-processing step which normalizes some miss-written characters to one base same form. It is sometimes simply refered to as "a normalization process" Ref. [6] Ref. [18]. Depending on the characters, there are two normlaization forms: reduced normalization(RED) and enriched normalization(ENR). Reduced normalization converts all Hamzated Alif(آ ، أ ، إ) into bare Alif (ا), and turns Alif Maqsura or the dotless Ya(ى) into a dotted Ya (ي), while enriched normalization chooses the appropriate form of Alif. The two forms were introduced in Ref. [18] by El Kholy and Habash. Badr et al. Ref. [6] experimented

[35] http://www.cs.umd.edu/ snover/tercom/

[36] http://www.cs.cmu.edu/ alavie/METEOR/

[37] http://www.kecl.ntt.co.jp/icl/lirg/ribes/

[38] https://code.google.com/archive/p/costa-mt-evaluation-tool/

with the reduced form only. Linguistically, the two forms change the meaning of some words and lead to non-correct Arabic text, but the enriched form of Arabic is more realistic and is desired to evaluate against.

2) Orthographic De-normalization

Orthographic De-normalization is a post-processing technique, to de-normalize normalized text. In order to produce correct Arabic script, a reduced tokenized (a morphological process that will be discussed in the next sub-section) output should be enriched and de-tokenized. Two methods were proposed by El Kholy and Habash in Ref. [17].

Normalizing text to the reduced form can be done through a simple characters substitution script, but converting it to an enriched form requires a machine learning algorithm. As stated above the enriched form can be produced using MADA toolkit Ref. [29]. There is a problem downloading MADA at the moment of writing this paper, but MADA was integrated with AMIRA Ref. [12], a toolkit for Arabic processing in 2014 with a title MADAMIRA Ref. [43]. MADAMIRA is free to use for research purposes[39].

Orthographic normalization is not the first step to pre-process Arabic text; a cleaning step is advised to be the best start. SPLIT is a unified preprocessing tool for SMT corpora, its goal is to standardize the preprocessing steps to avoid the drastic changes that are lead by various preprocessing techniques. SPLIT has normalization options beside the cleaning steps. SPLIT was developed by the Natural Language Processing research lab in George Washington University[40]. ALBadrashiny et al. described in Ref. [1] the details of SPLIT.

B. Morphologically Processing Techniques

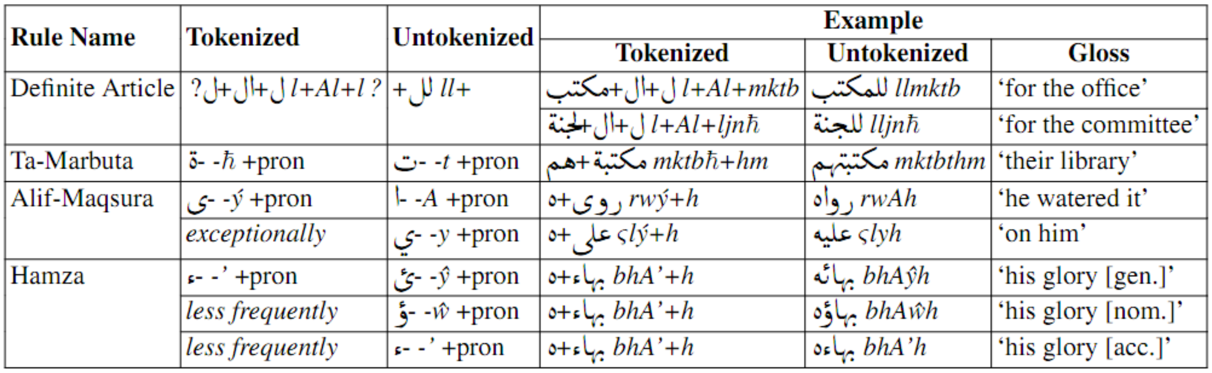
1) Morphological Tokenization

Morphological Tokenization is a pre-processing technique to separate the cliticized Arabic words into parts. Arabic words are highly cliticized (i.e. a word can have many part of speech tags (POS)), for example the word *wsnkAtbhum* وسنكاتبهم (which means "and we will write to them") is cliticized as following:

w+	s+	n+	kAtb+ hum
and+	will+	we+	write+ to them

Tokenization reduces sparsity on the Arabic side of the parallel corpus, and without tokenization, Arabic will have more surface forms than on the English side Ref. [6].

The terms "morphological tokenization" and "segmentation" are often used interchangeably Ref. [6] Ref. [18], despite a claim that there is a difference between the two terms. El Kholy and Habash Ref. [18] illustrated the difference with an example: segmentation of maktbthom (their library) - مكتبتهم is segmented as (maktbt + hum - مكتبت + هم) and this is not the right Arabic word (maktaba - مكتبة) . Some of the adjustment rules in the tokenization process according to El Kholy and Habash in Ref. [18], are illustrated in Figure 4.



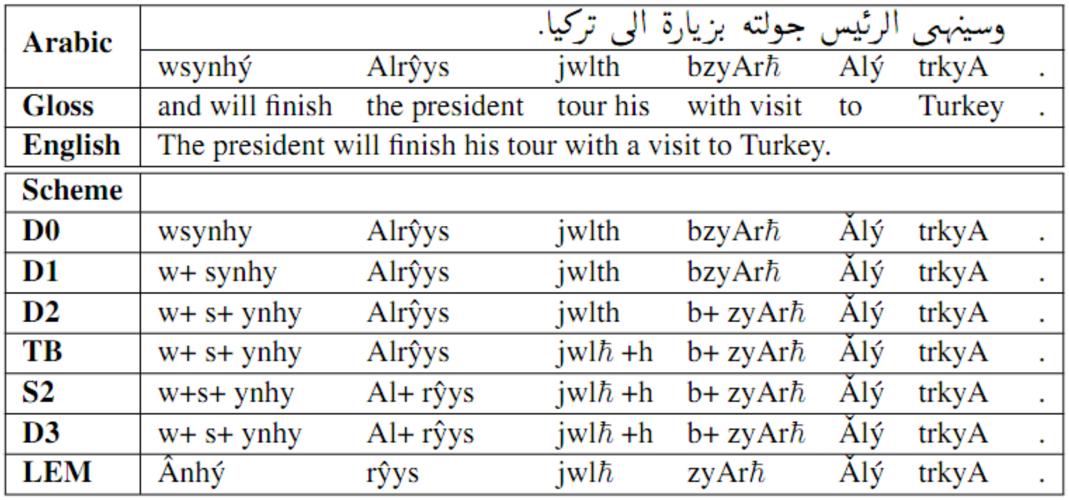
Rule Name	Tokenized	Untokenized	Example		
			Tokenized	Untokenized	Gloss
Definite Article	ل+ال+ل? *l+Al+l ?*	+لل *ll+*	ل+ال+مكتب *l+Al+mktb*	للمكتب *llmktb*	'for the office'
			ل+ال+لجنة *l+Al+ljnħ*	للجنة *lljnħ*	'for the committee'
Ta-Marbuta	ـة *-ħ* +pron	ـت *-t* +pron	مكتبة+هم *mktbħ+hm*	مكتبتهم *mktbthm*	'their library'
Alif-Maqsura	ـى *-ý* +pron	ـا *-A* +pron	روى+ه *rwý+h*	رواه *rwAh*	'he watered it'
	exceptionally	ـي *-y* +pron	على+ه *ςlý+h*	عليه *ςlyh*	'on him'
Hamza	ـء *-'* +pron	ـئ *-ŷ* +pron	بهاء+ه *bhA'+h*	بهائه *bhAŷh*	'his glory [gen.]'
	less frequently	ـؤ *-ŵ* +pron	بهاء+ه *bhA'+h*	بهاؤه *bhAŵh*	'his glory [nom.]'
	less frequently	ـء *-'* +pron	بهاء+ه *bhA'+h*	بهاءه *bhA'h*	'his glory [acc.]'

Figure 4 Examples of Arabic morphological adjustment rules. Source: Ref. [18]

[39] http://innovation.columbia.edu/technologies/

[40] http://care4lang1.seas.gwu.edu/split.php

Six schemes for Arabic morphological tokenization, were introduced by El Kholy and Habash Ref. [18]. The schemes are named D0, D1, D2, TB, S2, and D3 and illustrated with a sentence example in Figure5 (Note that: D0 is the word surface form). The six schemes build on the work of Badr et al. Ref. [6], with two enhancements: a comparison between the wide range of schemes (Badr et al. only used S2 and D3), and discussing the issue of producing unnormalized Arabic text (Badr et al. only experimented with normalized text). Moreover, El Kholy and Habash Ref. [18] mentioned that the results of S2 and D3 schemes have consistency with Badr et al. Ref. [6] results . They also noted that TB outperforms S2 scheme , and experiments with the reduced Arabic text then enriching it occasionally produce results better than training with enriched text directly.



Arabic	وسينهى الرئيس جولته بزيارة الى تركيا.						
	wsynhý	Alrŷys	jwlth	bzyArħ	Alý	trkyA	.
Gloss	and will finish	the president	tour his	with visit	to	Turkey	.
English	The president will finish his tour with a visit to Turkey.						
Scheme							
D0	wsynhy	Alrŷys	jwlth	bzyArħ	Ǎlý	trkyA	.
D1	w+ synhy	Alrŷys	jwlth	bzyArħ	Ǎlý	trkyA	.
D2	w+ s+ ynhy	Alrŷys	jwlth	b+ zyArħ	Ǎlý	trkyA	.
TB	w+ s+ ynhy	Alrŷys	jwlħ +h	b+ zyArħ	Ǎlý	trkyA	.
S2	w+s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ǎlý	trkyA	.
D3	w+ s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ǎlý	trkyA	.
LEM	Ânhý	rŷys	jwlħ	zyArħ	Ǎlý	trkyA	.

Figure 5 A sentence in the various tokenization schemes. Source: Ref. [18]

2) *Morphological De-tokenization*

Morphological Detokenization is a post-processing technique used to convert tokenized Arabic output to its original, uncliticized form. This process is called morphological detokenization and recombination interchangeably Ref. [17] Ref. [6]. Four techniques for detokenization were introduced in Ref. [6] by Badr et al.: (S)Simple, (R)Rule-based, (T)Table-based and (T+R)Table+Rule. El Kholy and Habash Ref. [17] added two techniques: (T+LM)Table+Language Modeling, and (T+R+LM). The work presnted in Ref. [6] reported that (T+R) was the best technique, while the work presented in Ref. [17] reported that (T+R+LM), one of the two techniques added later, was the best.

There are processing tools for Arabic morphological tokenization/detokenization, often called segmenters. The Stanford Arabic segmenter was released in 2014, and it is an implementation of the segmenter detailed in Ref. [38]. The Stanford segmenter is available for research purposes[41]. It is also a tool for orthographic normalization. In 2014, MADAMIRA was released Ref. [43]. MADAMIRA is a system for processing Arabic that has an Arabic morphological analysis and disambiguation module, in addition to a segmentation module. In 2016, QCRI (Qatar Computing Research Institute) released the FARASA segmenter Ref. [4]. FARASA has other modules for Arabic processing inlcuding a POS tagger, a diacritizer, and a dependency parser.

3) *Factored Models*

Factored Models are used to make training more reliable. In 2013, Khemakhem et al. Ref. [31] highlighted a problem in MT scoring. The problem, as they mentioned, is that MT scoring relies on the words history more than other features of the words. For example, the word *katab* means (to write) and the word *kotob* means (books), have the same surface form but different meaning according to context. Here the diacritics of the Arabic words should be important features in training the parallel corpus. Khemakhem et al. proposed two features for Arabic words which are the the word and its syntactic class (e.g. noun, verb, particle and proper noun).

Earlier in 2008, Badr et al. Ref. [6] experimented with factored models. The factors were on the two sides of the corpus (English and Arabic). English factors were: the surface form and POS tag, and the Arabic factors were: the

[41] http://nlp.stanford.edu/software/segmenter.shtml

surface form, stem, and the POS tag along with the segmented clitics; for example, the Arabic word *wlAwlAdh* means (and for his kids) has the following factors: *AwlAd* and *w+l+N+P:3MS*.

C. Syntactically Processing Techniques

1) Syntactic Reordering

Syntactic Reordering is a process that aims at overcoming the linguistic gap between Arabic syntax and English syntax. English sentences are written in SVO order, while Arabic text favors the VSO order. Badr et al. Ref. [7] applied a set of rules on the source language (i.e.English) for a better alignment. Using a parse tree the rules are:

1. NP(Noun Phrase): inverts all nouns, adjectives and adverbs into a NP.
2. PP(Prepositional Phrase): transforms prepositional phrases of the form N1 of N2 … of Nn into N1 N2 … Nn.
3. Definite article (the): replicates "the" before adjectives.
4. VP(Verb Phrases): converts SVO order into VSO.

Badr et al. Ref. [7] used the Collins parser Ref. [11] for the English corpus, after tagging it with the Stanford POS tagger, and splitting the text into smaller sentences. After that they tagged them using a maximum entropy tagger Ref. [48]. A named entity recognition(NER) step was carried out by the Stanford NER for location, person, and organization entities. They dicovered that the rule that replicates the "the" hurts the translation score.

On the other hand, Habash Ref. [28] experimented similar rules for the opposite direction (i.e. Ar-En SMT). The results were less promising compared with En-Ar direction. Arabic parsers have less quality than English parsers, this might be the cause. But few Arabic parsers were released after this publication such as: Stanford Arabic parser Ref. [26], and it is needed to be tested with an SMT task.

D. Multi-word Expressions

In the scope of machine translation, multi-word expressions are the phrases or sentences that have different meaning than the literal meaning for each word separately. Detecting MWEs is also an active area of research. It usually poses an issue for junior human translators who are not professional in the source language. A significant amount of research was done for detecting multi-word expressions, but the amount of research concerning integrating MWEs into MT systems is not significant.

Eventhough modeling MWEs in SMT is a hard task, Ghoneim and Diab Ref. [22] described three methods to integrate MWEs into the Moses SMT system. Their work was an extension of Carpuat and Diab Ref. [9]. The study concentrated on how the integration methods are done, and focused less on how the MWE extraction process happens. MWEs were extracted from lexical databases, the English WordNet 3.0, and using named entity recognizers (NEs are a type of MWE). In the scope of detecting MWEs, an MWEtoolkit can help in this task; because an MWEtoolkit is a framework for language-independent MWE identification from corpora Ref. [47]. Moreover, Attia et al. published an automatic technique to extract Arabic MWEs Ref. [5].One of the coauthors of this work, an Arabic linguist, published a manually extracted Arabic MWEs on his personal website[42].

CONCLUSIONS

SMT is an active research area and it is needed to have concerned researchers, junior scientists and institutions to enhance En-Ar direction. Despite that En-Ar is less represented in SMT research community, the linguistic enhancements are promising for more improvements. Recently, SMT Arabic research community concerns with Ar-En dialectally (i.e. Egyptian Arabic-En, Syrian Arabic-En ,... etc), but we encourage working with Standard Arabic SMT instead of dialect; because the news, books, and science documents are needed to be understandable by all Arabic people not a single country.

[42] http://www.attiaspace.com/

REFERENCES

[1] M. Al-Badrashiny, A. Pasha, M. Diab, N. Habash, O. Rambow, W. Salloum, and R. Eskander, "Split: Smart preprocessing (quasi) language independent tool," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (N. C. C. Chair), K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Paris, France), European Language Resources Association (ELRA), may 2016.

[2] S. Green, D. Cer, and C. D. Manning, "Phrasal: A toolkit for new directions in statistical machine translation," in *In Proceddings of the Ninth Workshop on Statistical Machine Translation*, 2014.

[3] C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre, "Docent: A document-level decoder for phrase-based statistical machine translation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Sofia, Bulgaria), pp. 193–198, Association for Computational Linguistics, August 2013.

[4] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 11–16, Association for Computational Linguistics, San Diego, California, 2016.

[5] M. Attia, L. Tounsi, P. Pecina, J. van Genabith, and A. Toral, "Automatic extraction of arabic multiword expressions," 2010.

[6] I. Badr, R. Zbib, and J. Glass, "Segmentation for english-to-arabic statistical machine translation," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pp. 153–156, Association for Computational Linguistics, 2008.

[7] I. Badr, R. Zbib, and J. Glass, "Syntactic phrase reordering for english-to-arabic statistical machine translation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 86–93, Association for Computational Linguistics, 2009.

[8] J. Cadigan and Y. Marton, "Gisa: Giza++ implementation over spark by apache," *AMTA 2016*, p. 13, 2016.

[9] M. Carpuat and M. Diab, "Task-based evaluation of multiword expressions: a pilot study in statistical machine translation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 242–245, Association for Computational Linguistics, 2010.

[10] K. Chatzitheodorou, "Costa mt evaluation tool: An open toolkit for human machine translation evaluation," *The Prague Bulletin of Mathematical Linguistics*, vol. 100, pp. 83–89, 2013.

[11] M. Collins, "Three generative, lexicalised models for statistical parsing," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 16–23, Association for Computational Linguistics, 1997.

[12] M. Diab, "Second generation amira tools for arabic processing: Fast and robust tokenization, pos tagging, and base phrase chunking," in *2nd International Conference on Arabic Language Resources and Tools*, 2009.

[13] N. Durrani, H. Sajjad, H. Hoang, and P. Koehn, "Integrating an unsupervised transliteration model into statistical machine translation.," in *EACL*, vol. 14, pp. 148–153, 2014.

[14] C. Dyer, J. Weese, H. Setiawan, A. Lopez, F. Ture, V. Eidelman, J. Ganitkevitch, P. Blunsom, and P. Resnik, "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models," in *Proceedings of the ACL 2010 System Demonstrations*, pp. 7–12, Association for Computational Linguistics, 2010.

[15] S. Ebrahim, D. Hegazy, M. G. Mostafa, and S. R. El-Beltagy, "English-arabic statistical machine translation: State of the art," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 520–533, Springer, 2015.

[16] A. Eisele and Y. Chen, "Multiun: A multilingual corpus from united nation documents.," in *LREC*, 2010.

[17] A. El Kholy and N. Habash, "Techniques for arabic morphological detokenization and orthographic denormalization," in *Editors & Workshop Chairs*, p. 45, 2010.

[18] A. El Kholy and N. Habash, "Orthographic and morphological processing for english–arabic statistical machine translation," *Machine Translation*, vol. 26, no. 1-2, pp. 25–45, 2012.

[19] I. A. El-Khair, "Abu el-khair corpus: A modern standard arabic corpus,"

[20] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, p. 14, 2009.

[21] C. Federmann, "Appraise: an open-source toolkit for manual evaluation of mt output," *The Prague Bulletin of Mathematical Linguistics*, vol. 98, pp. 25–35, 2012.

[22] M. Ghoneim and M. T. Diab, "Multiword expressions in the context of statistical machine translation.," in *IJCNLP*, pp. 1181–1187, 2013.

[23] J. González, G. Sanchis, and F. Casacuberta, "Learning finite state transducers using bilingual phrases," in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 411–422, Springer, 2008.

[24] J. González and F. Casacuberta, "Great: a finite-state machine translation toolkit implementing a grammatical inference approach for transducer inference (giati)," in *Proceedings of the EACL 2009 Workshop on Computational Linguistic Aspects of Grammatical Inference*, pp. 24–32, Association for Computational Linguistics, 2009.

[25] J. González and F. Casacuberta, "Great: open source software for statistical machine translation," *Machine translation*, vol. 25, no. 2, pp. 145–160, 2011.

[26] S. Green and C. D. Manning, "Better arabic parsing: Baselines, evaluations, and analysis," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 394–402, Association for Computational Linguistics, 2010.

[27] N. Habash, A. Soudi, and T. Buckwalter, "On arabic transliteration," in *Arabic computational morphology*, pp. 15–22, Springer, 2007.

[28] N. Habash, "Syntactic preprocessing for statistical machine translation," *MT Summit XI*, pp. 215–222, 2007.

[29] N. Habash, O. Rambow, and R. Roth, "Mada+ tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization," in *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, pp. 102–109, 2009.

[30] K. Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Association for Computational Linguistics, 2011.

[31] I. T. Khemakhem and S. Jamoussi, "Integrating morpho-syntactic features in english-arabic statistical machine translation," *ACL 2013*, p. 74, 2013.

[32] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," *Machine translation: From real users to research*, pp. 115–124, 2004.

[33] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT summit*, vol. 5, pp. 79–86, 2005.

[34] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177–180, Association for Computational Linguistics, 2007.

[35] P. Koehn, "An experimental management system," *The Prague Bulletin of Mathematical Linguistics*, vol. 94, pp. 87–96, 2010.

[36] A. Lardilleux and Y. Lepage, "Sampling-based multilingual alignment," in *Recent Advances in Natural Language Processing*, pp. 214–218, 2009.

[37] J. B. Marino, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R. Costa-Jussà, "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.

[38] W. Monroe, S. Green, and C. D. Manning, "Word segmentation of informal arabic with domain adaptation.," in *ACL (2)*, pp. 206–211, 2014.

[39] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[40] M. Olteanu, C. Davis, I. Volosen, and D. Moldovan, "Phramer: an open source statistical phrase-based translator," in *Proceedings of the Workshop on Statistical Machine Translation*, pp. 146–149, Association for Computational Linguistics, 2006.

[41] D. Ortiz-Martnez and F. Casacuberta, "The new thot toolkit for fully-automatic and interactive statistical machine translation," in *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, pp. 45–48, 2014.

[42] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311–318, Association for Computational Linguistics, 2002.

[43] A. Pasha, M. Al-Badrashiny, A. E. Kholy, R. Eskander, M. Diab, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," in *In Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*, 2014.

[44] B. Paul, B. Phil, and H. Hieu, "Oxlm: A neural language modelling framework for machine translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 102, no. 1, pp. 81–92, 2014.

[45] A. Pauls and D. Klein, "Faster and smaller n-gram language models," in *Proceedings of ACL*, (Portland, Oregon), Association for Computational Linguistics, June 2011.

[46] A. Rafalovitch, R. Dale, *et al.*, "United nations general assembly resolutions: A six-language parallel corpus," in *Proceedings of the MT Summit*, vol. 12, pp. 292–299, 2009.

[47] C. Ramisch, A. Villavicencio, and C. Boitet, "mwetoolkit: a framework for multiword expression identification.," in *LREC*, vol. 10, pp. 662–669, 2010.

[48] A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," in *Proceedings of the conference on empirical methods in natural language processing*, vol. 1, pp. 133–142, Philadelphia, PA, 1996.

[49] A. Stolcke *et al.*, "Srilm-an extensible language modeling toolkit.," in *Interspeech*, vol. 2002, p. 2002, 2002.

[50] D. Talbot and M. Osborne, "Randomised language modelling for statistical machine translation," in *ACL*, vol. 7, pp. 512–519, 2007.

[51] S. M. Translation, "Amt," 2007.

[52] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, "Decoding with large-scale neural language models improves translation.," in *EMNLP*, pp. 1387–1392, Citeseer, 2013.

[53] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, "The united nations parallel corpus v1. 0,"

BIOGRAPHY

Sara Ebrahim is a Teaching Assistant at the Scientific Computing Department, Faculty of Computer and Information Sciences (FCIS), Ain Shams University, Cairo, Egypt. She received her B.Sc. degree from the Scientific Computing Department, FCIS in 2010. She is about to finish her M. Sc. degree in July 2017. She also received a professional diploma in media and literary translation from the School of Continuing Education, American University in Cairo (SCE AUC) in 2016. She is in her final year of studying the Arabic and Islamic studies in Dar El-Olum Faculty, Open Education Center, Cairo University, Giza, Egypt. Sara's main research interests focus on Arabic Natural Language Processing, Information Retrieval, Statistical Machine Translation and Social Media Analysis.

Dr. Samhaa R. El-Beltagy: is a currently a full Professor at Nile University (NU), Center for Informatics Science, where she's director of the Informatics Program and head of the Text Mining Research Group. She received her PhD in Computer Science from the University of Southampton, UK in 2001, and her Masters and Bachelor degrees in Computer Science from the American University in Cairo in 1997 and 1993 respectively. After completing her PhD, Dr. El-Beltagy has taken on technical leadership of numerous national developmental projects. In 2009, she was awarded the title of ACM senior member. Over the past 10 years, Dr. El-Beltagy been focused on the area of text analytics. During the past two years, she has given three keynote speeches in international conferences about Social Media Analytics. She currently has over seventy refreed international publications and has served and continues to serve on the international program committees of numerous reputable international conferences and workshops, directly and indirectly related to the general field of "Data Analytics". She has also served as an external reviewer for a number of international journals, and national projects.

Dr. Doaa Hegazy is an Associate Professor at the Scientific Computing Department, Faculty of Computer and Information Sciences (FCIS), Ain Shams University, Cairo, Egypt. She received her B. Sc. and M. Sc. degree from the Scientific Computing Department, FCIS in 2000 and 2004 respectively and she received her Ph.D. degree (Dr.rer.nat.) from faculty of Mathematics and Computer Science, Friedrich Schiller University in Jena, Germany in 2009 (through A DAAD scholarship.) The main research interests of Dr. Doaa Hegazy focus on Image Processing, Computer Vision, AI, Computer Graphics, Scientific Visualization, and Augmented Reality.

Mostafa G. M. Mostafa is a Professor (full, 2007) of Computer Science at the Faculty of Computer and Information Sciences (FCIS), Ain Shams University. He served as a CS Department Chair and a Vice-Dean at FCIS. He received a B.Sc. (Honor) in 1984, a M.Sc. in 1989 and a Ph.D. in 1996, in Computational Physics from the Faculty of Science, Ain Shams University, Cairo, Egypt. He worked as a research scientist at Physics Division, Oak Ridge National Lab (ORNL), USA, as a Postdoc at the Department of Electrical and Computer Engineering, University of Louisville, USA, and as a Professor at the Faculty of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia. He published more than 60 scientific articles in international and local journals and conferences. His research interests include: Machine Learning, Pattern Recognition, Computer Vision, Medical Image Analysis, Arabic OCR, Data Mining, Bioinformatics, and Information Security.

نحو تطوير نظام للترجمة الآلية من اللغة الإنجليزية إلى اللغة العربية بالطرق الإحصائية المعتمدة على ترجمة الجُمل

سارة إبراهيم[1*]، سمحاء البلتاجي [2**]، دعاء حجازي [3*]، مصطفى جادالحق[4***]

*قسم الحسابات العلمية، كلية الحاسبات والمعلومات، جامعة عين شمس، القاهرة، مصر
[1]sara.elkafrawy@gmail.com
[3]doaa.hegazy@cis.asu.edu.eg

** مركز علوم الحاسب، جامعة النيل، مدينة السادس من أكتوبر، مصر
[2]samhaa@computer.org

***قسم علوم الحاسب، كلية الحاسبات والمعلومات، جامعة عين شمس، القاهرة، مصر
[4]mgmostafa@cis.asu.edu.eg

ملخص

يتناول هذا البحث شرح خطوات إجراء الترجمة الآلية بالطرق الإحصائية المعتمدة على الجُمل وذلك من اللغة الإنجليزية إلى اللغة العربية. كما يعرض البحث أهم التجارب التي أجريت لتحسين الترجمة الآلية التي تستهدف الترجمة إلى اللغة العربية. ويركز البحث على إظهار جميع المدونات المترجمة المتاحة لإجراء مثل هذه التجارب، وعرض الأفكار اللغوية التي تهدف لتحسين الترجمة الآلية بالطرق الإحصائية المبنية على ترجمة الجُمل. وجدير بالذكر أن هدف هذا البحث هو مساعدة من يريد بناء نظامٍ ضخم للترجمة الآلية من اللغة الإنجليزية إلى اللغة العربية، هذا بالإضافة إلى أن البحث يوضح أهم التحديات اللغوية العامة التي تواجه الترجمة الآلية للغةٍ العربية. فيمكن أن نعتبر هذا البحث مرشدًا لبناء أنظمة الترجمة الآلية المعتمدة على ترجمة الجُمل بالطرق الإحصائية من اللغة الإنجليزية إلى اللغة العربية، بالرغم من أن الخطوات المشروحة تنطبق على أي زوجين من اللغات.