Evaluation of Arabic Machine Translation System based on the Universal Networking Language

Noha Adly[1,2] Sameh Al Ansary[1,3]

[1]Bibliotheca Alexandrina, Alexandria, Egypt
[2]Department of Computer and Systems Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt
[3]Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt

Abstract. This paper evaluates a machine translation (MT) system based on the interlingua approach, the Universal Network Language (UNL) system, designed for Multilanguage translation. The study addresses evaluation of English-Arabic translation and aims at comparing the MT systems based on UNL against other systems. Also, it serves to analyze the development of the system understudy by comparing output at the sentence level. The evaluation is performed on the Encyclopedia of Life Support Systems (EOLSS), a wide range corpus covering multiple linguistic and cultural backgrounds. Three automated metrics are evaluated, namely BLEU, F_1 and F_{mean} after being adapted to the Arabic language. Results revealed that the UNL MT outperforms other systems for all metrics.

Keywords: Machine Translation, Natural Language Processing, Natural Language Generation, Evaluation of MT, Universal Networking Language, Encyclopedia of Life Support Systems, Interlingua.

1 Introduction

Research in machine translation (MT) has spanned several approaches. Statistical machine translation has been the approach most widely used, see [13] for a recent survey. The Interlingua approach relies on transforming the source language to a language-independent representation, which can then be transformed to the target language. When multilingual translation is of interest, the interlingua approach allows to build a system of N languages with a linear effort while the statistical approach would require a quadratic effort. The challenge with the interlingua approach is to design a language independent intermediate representation that captures the semantic structures of all languages while being unambiguous. The interlingua has been used on limited task-oriented domains such as speech translation for specific domains [8]. Few efforts studied machine translation based on Interlingua, but on a limited scale, for Indian languages [20], Korean language [10] and Arabic language [19].

The UNL System promises a representation of all information, data and knowledge that humans produce in their own natural languages, in a language independent way, with the purpose of overcoming the linguistic barrier in Internet. The UNL is an

artificial language that has lexical, syntactical and semantic components as does any natural language. This language has been proven tractable by computer systems, since it can be automatically transformed into any natural language by means of linguistic generation processes. It provides a suitable environment for computational linguists to formalize linguistic rules initiation from semantic layer.

The Encyclopedia of Life Support Systems (EOLSS) [6] is an Encyclopedia made of a collection of 20 online encyclopedias. It is a massive collection of documentation, under constant change, aiming at different categories of readers coming from multiple linguistic and cultural backgrounds. EOLSS is an unprecedented global effort over the last ten years, with contributions from more than 6000 scholars from over 100 countries, and edited by nearly 300 subject experts. The result is a virtual library equivalent to 200 volumes, or about 123,000 printed pages.

Availing EOLSS in multiple languages is a main goal of its initiators. However, translating EOLSS in every possible language is a daunting task that requires years of work and large amount of human and financial resources, if done in the conventional ways of translation. The UNDL Foundation proposed to use the UNL System for representing the content of EOLSS in terms of language independent semantic graphs, which in turn can be decoded into a target natural language, generating a translation of EOLSS documents into multiple languages. With the UNL System, this can be achieved in a relative shorter period of time, and at lower costs in comparison to costs of traditional translation. Work has actually started with the six official languages of the United Nations. 25 documents, forming around 15,000 sentences have been enconverted from EOLSS to UNL. The UNL version of EOLSS is sent to the UNL language centers for deconversion. It is a prototype for translating massive amount of text; done in anticipation to the deconversion in many other languages of the world.

The Arabic language center has completed the deconversion of the 25 documents of the prototype and automatically generated the equivalent Arabic language text. The purpose of this paper is to evaluate the quality of the translated text. The objective of the evaluation is twofold. First, it is desirable to evaluate the strength and weakness of the machine translation generated through the UNL system and compare it against other MT systems. Second, it is aimed to set up a framework of evaluation that can be applied on a frequent and ongoing basis during the system development, in order to guide the development of the system based on concrete performance improvements.

The rest of the paper is organized as follows. Section 2 gives a brief description of the UNL system and describes its usage for the automated translation of the EOLSS. Section 3 presents a brief description of the Arabic dictionary and generation rules. Section 4 describes the automated metrics used in the performance evaluation and introduces some adaptation of the metrics to suit the Arabic language. Section 5 gives an overview of the process of the data preparation and presents the experimental design. Section 6 presents the different conducted experiments and discusses the results. Finally, Section 7 concludes the paper.

2 The UNL System

The architecture of the UNL system (Fig. 1) comprises three sets of components [23]:

1. *Linguistic components*: dictionaries that include Universal Words (UWs) and their equivalents in natural languages, grammatical rules responsible for producing a well formed sentence in the target natural language and knowledge base for representing a universal hierarchy of concepts in natural languages;
2. *Software components*: two software programs for converting content from natural languages to UNL (the EnConverter) and vice versa (the DeConverter). The EnConverter is a language independent parser that provides synchronously a framework for morphological, syntactic and semantic analysis. It is designed to achieve the task of transferring the natural language to the UNL format or UNL expressions. The DeConverter is a language independent generator that provides synchronously a framework for morphological and syntactic generation, and word selection for natural collocation. DeConverter can deconvert UNL expressions into a variety of native languages, using the Word Dictionary, formalized linguistic rules and Co-occurrence Dictionary of each language;
3. *System interfacing components*: protocols and tools enabling the flow of UNL documents throughout the web.

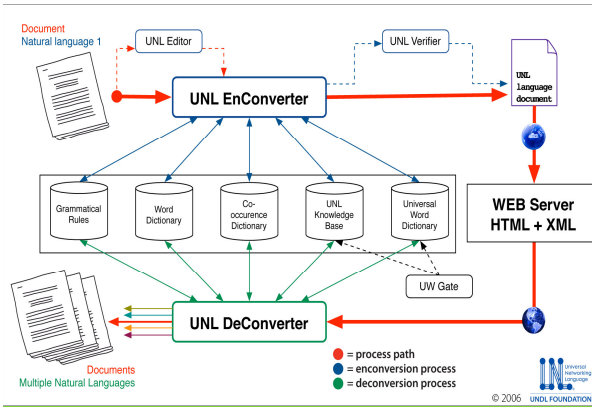


Fig. 1: The core architecture of the UNL system

2.1. UNL Language Components

The UNL consists of Universal Words (*UW*s), *Relations*, *Attributes*, and UNL Knowledge Base (UNL KB). The *UW*s constitute the vocabulary of the UNL, *Relations* and *Attributes* constitute the syntax of the UNL and the UNL KB constitutes the semantics of the UNL. Formally, a UNL expression can be viewed as a semantic network, whose nodes are the *UW*s, linked by arcs labeled with the UNL *Relations* which express the objective meaning of the speaker. *UW*s are modified by the so-called *Attributes* to convey the subjective meaning of the speaker (For more details see [23]). The UNL KB constitutes the semantic background of the UNL System. It is constituted by the binary direct relations between two UWs. With these

links, a conceptual network can be shaped to form a lattice structure. The structure allows for implementing the principle of inheritance in the definition of concepts.

The UNL KB is meant to assure robustness and precision to the UNL System, both to the NL-UNL encoverting, and to the UNL-NL deconverting processes. In the former case, the UNL KB would be used as a sort of word sense disambiguation device. In the latter, the UNL KB would allow for the deconversion of UWs not enclosed in the target language dictionaries.

2.2. Using UNL in Machine Translation of EOLSS

Translation with the UNL system is a two-step process. The first step deals with Enconverting the content of the EOLSS from the source language (English) to UNL (the universal representation). This process is called the *UNLization* process; it is carried out with the use of the English-UNL Enconverter. Initially, some post-editing is needed, but as the performance of the English Enconverter and the technical dictionaries improve, human intervention will be gradually reduced, and productivity will be increased.

The second step deals with Deconverting EOLSS content from UNL to a target natural language [2, 3]. This Deconversion process is a task to be carried out by the UNL-Language Server of each language. Each UNL Language Server contains a dictionary and generation rules (deconversion), working in association with the UNL KB, which are the enabling components in this process. Since we are concerned with the generation of the Arabic language, we briefly describe the design of the Arabic dictionary and generation rules in the next section.

3 Generating Arabic from the UNL Interlingua

3.1. Design of the UNL-Arabic Dictionary

The Arabic dictionary is designed to support morphological, syntactic and semantic analysis and generation needed for both Arabic EnConversion and DeConversion rules. The design of the dictionary includes the Arabic word heading, its corresponding meanings, and information on its linguistic behavior. The focus of attention is given to the form of the head word of the entry needed to fulfill language analysis and generation tasks adequately. The entries are stem-based to avoid adding all possible inflectional and derivational paradigms of each lexical item to the dictionary, and to minimize the number of entries in the dictionary which will give more efficiency in the analysis and generation tasks and minimize the processing time (e.g. instead of storing أكاديمية، أكاديميتنا، أكاديميات، أكاديمي etc., only أكاديمي will be stored). The Arabic UNL dictionary stores three types of linguistic information. First, morphological information which is responsible for correctness of the morphology of words; it describes the changes that occur within a word when it is attached to various suffixes and prefixes in different contexts. Second, syntactic information to generate well-formed Arabic sentence structure; it determines grammatical relations coded as

the presence of adjuncts and arguments in isolation or as sub-categorization frames, and describes grammatical relations between words. Third, semantic information about the semantic classification of words that allows for correct mapping between semantic information in UNL-graphs and syntactic structure of the sentence under generation. The following examples represent full records of lexical entries:

[أوى]{أوَى}"accommodate(agt>person,obj>person)"(ST,1.2,2V,3V,V1,1?) <A,0,0>;
[ئو] {ائْو} "accommodate(agt>person,obj>person)" (ST,1.2,2V,3V,V6,1?) <A,0,0>;
[أوي]{أوَيْتُ}"accommodate(agt>person,obj>person)"(ST,1.2,2V,3V,V2,1?)<A,0,0>;
[أو] {أو} "accommodate(agt>person,obj>person)" (ST,1.2,2V,3V,V5,1?) <A,0,0>;
[آوي]{آوي}"accommodate(agt>person,obj>person)"(ST,1.2,2V,3V,V3,1?) <A,0,0>;
[ؤوي]{يؤوي}"accommodate(agt>person,obj>person)"(ST,1.2,2V,3V,1?) <A,0,0>;
[آو] {آو} "accommodate(agt>person,obj>person)" (ST,1.2,2V,3V,V4,1?) <A,0,0>;
[اوا]{أوَاه}"accommodate(agt>person,obj>person)" (ST,1.2,2V,3V,V7,1?) <A,0,0>;

The example above shows the different word forms of the verb "أوى" that are stored in the Arabic dictionary with different linguistic information about each form to guide the grammar to pick the appropriate one according to the syntactic structure and the tense of the sentence.

3.2. Design of the Arabic Generation Grammar

The Arabic language is a morphologically and syntactically rich language and its generation is very complex. Hence, the technical design of the Arabic generation grammar is divided into several stages, namely the lexical mapping stage, the syntactic stage and the morphological stage. The lexical mapping stage deals with identifying the target lexical items. The syntactic stage deals with the order of words in the node list, and morphological stage specifies how to form words and deals with agreement gender, number, person and definiteness. The different stages are illustrated in Fig. 2.

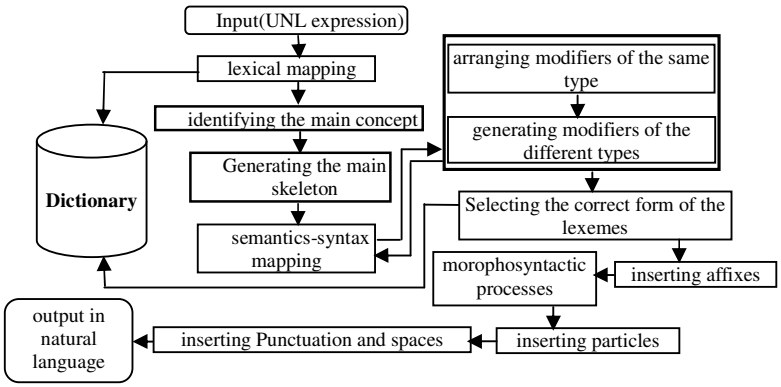


Fig. 2. A block diagram describing Arabic generation from interlingua

Lexical Mapping: The lexical mapping stage performs the mapping between the meaning conveyed by the concepts of the intermediate representation (UNL intelingua) and the lexical items of the target language. For example, the word "answer" can be translated in the Arabic language as "يجيب" or "إجابة" but it is expressed by two concepts "answer(agt>thing,obj>thing)" which is mapped with the corresponding Arabic verb "يجيب" and the concept "answer(icl>notion) which is mapped with the corresponding Arabic noun"إجابة".

Syntactic Stage: The syntactic stage is concerned with the order of words in the node list; it can be divided into two phases. The first phase is concerned with building the main skeleton of the sentence. The starting node in the UNL network is the 'entry' node that refers to the main concept of the sentence which is marked as "@entry". The phase continues to generate the arguments of the main concept concerning the suitable Arabic syntactic structure in either a nominal structure (Topic-Comment) or in a verbal structure (VSO). The second phase in the grammar deals with generating the modifiers. One of the challenges faced in this stage is when a given node in the semantic network is modified by more than one modifier of the same type. The Arabic generation grammar is designed to control the insertion of nodes in such a situation. The generation process highlights a basic point which is the type and number of syntactic arguments a predicate takes are determined by the type and number of semantics arguments that a predicate expresses. This actually reflects the interface between semantics and syntax in Natural Language Generation.

Morphological stage: The Morphological stage is concerned with two axes. First, inserting affixes (prefixes and suffixes) to the node list to generate the final form of the entries according to the linguistic features attached to each entry in the dictionary. The features are in turn based on the form of the dictionary entries selected to represent different paradigms representing lexemes. For example, the form of the defective verb "كان" 'be' changes according to subject pronouns. Therefore, three forms have been designed to represent all possible paradigms of this verb as shown in Table 1:

Table 1. The different paradigms of the same lexeme.

Hw	Reading	Uw	Pattern	V1	V2	V3	V_form
[كان]	كَانَ	be(aoj>thing,obj>thing)	1.1	Null	2V	Null	V1
[كن]	كن	be(aoj>thing,obj>thing)	1.1	Null	2V	Null	V3
[كون]	يكون	be(aoj>thing,obj>thing)	1.1	Null	2V	Null	V2

Each of the entries is given a different code, to be used in selecting the form required to represent the concept "be(aoj>thing,obj>thing)". In addition, based on the subject of the sentence a given affix will be added to the head word to generate the realized form. Second, inserting prepositions, attributes, pronouns that are needed because of the Arabic syntactic structure under generation and inserting punctuation marks. Spaces will be added at the end of the morphological phase after inserting all nodes from the node net. Spaces separate all nodes except nodes that represent affixes.

4 Performance Evaluation Metrics

Research in MT depends heavily on the evaluation of its results. Many automated measures have been proposed to facilitate fast and cheap evaluation of MT systems. Most efforts focus on devising metrics based on measuring the closeness of the output of MT systems to one or more human translation; the closer it is, the better it is. The challenge is to find a metric to be produced at low cost while correlating highly with human evaluation. The metric should be consistent and reliable. The most commonly used MT evaluation metric in recent years has been BLEU [15], an *n*-gram precision metric that demonstrated a high correlation with human judgment of system adequacy and fluency.

Various researchers have noted, however, some shortcomings in the metric due to being mainly a precision metric and its lack of consideration of the recall. Recall has been found to be extremely important for assessing the quality of MT output [9], as it reflects to what degree the candidate translation covers the entire content of the reference translation. Several metrics have been introduced recently that take precision and recall into account. GTM [14, 22] used a balanced harmonic mean of unigram precision and recall. METEOR [9] used a weighted harmonic mean placing more weight on recall than on precision and shown that this leads to better correlation. Recent development of METEOR [1, 4, 7] introduced unigram matching based on stemmed forms and synonyms matching using Wordnet. Other proposed methods for MT evaluation include TER [21], a metric based on the Levenshtein distance, but applied on the word level rather than the character level. It measures the number of edit operations needed to fix a candidate translation so that it semantically matches a reference translation. A related metric is CDER [11], which is based on the edit distance but accounts for an operation that allows for reordering of word blocks.

Several evaluations of the above metrics were conducted [12, 17] but there were no conclusions as to whether one of them supersedes the others. To achieve a balance in our evaluation, we chose BLEU, as it has been the primary metric used by most systems. But also we selected two metrics that incorporates recall, namely F_1 and F_{mean} which are based on GTM. These will be described in the following.

4.1. BLEU Metric

The main principle behind BLEU [15] is the measurement of the overlap in unigrams and higher order *n*-grams of words, between a *candidate* translation being evaluated and a set of one or more *reference* translations. The main component of BLEU is *n*-gram precision: the proportion of the matched *n*-grams out of the total number of *n*-grams in the candidate translation.

To avoid exceeding the counts of a word in the candidate with respect to its occurrence in any single reference, they introduced the *modified n-gram precision*. All candidate *n*-gram counts and their corresponding maximum reference counts are computed. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number (unclipped) of candidate *n*-grams. The precision p_n for each *n*-gram order is computed separately, and the precisions are combined via a geometric averaging.

Recall, which is the proportion of the matched *n*-grams out of the total number of *n*-grams in the reference translation, is not taken into account directly by BLEU. Instead, BLEU introduces a Brevity Penalty, which penalizes translations for being "too short". The brevity penalty is computed over the entire corpus and was chosen to be a decaying exponential in *r*/*c*, where *c* is the length of the candidate corpus and *r* is the effective length of the reference corpus. Therefore

$$\text{BLEU} = \text{BP}.\exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \text{ where BP} = \begin{cases} 1 & if\ c > r \\ e^{(1-\frac{r}{c})} & if\ c \leq r \end{cases}$$

The BLEU metric captures two aspects of translation: *adequacy* and *fluency*. Adequacy accounts for setting the words right, which is measured by BLEU with small *n*. Fluency captures the word order, the construction of the sentence and its well-formedness. It has been shown in [12,16] that shorter *n*-grams correlates better with adequacy with 1-gram being the best predictor, while longer *n*-grams has better fluency correlation. Typical values used by most systems is BLEU-4 [12]. The Smoothed technique proposed in [12] has been implemented in order to account for reliable score at the sentence level.

4.2 F_1 and F_{mean} Metrics

Both F_1 and F_{mean} metrics take into account Precision *P* and recall *R* and are based on unigram matching. F_1 is the harmonic mean[18] of the precision and recall, $F_1 = \frac{2PR}{P+R}$. F_{mean} [9], is similar to F_1, but recall is weighted more heavily than precision. $F_{mean}= \frac{10PR}{9P+R}$ where the weights are chosen to maximize the correlation with human judgment.

The definition of precision *P* and recall *R*, are adopted from [14,22]: given a set of candidates *Y* and a set of references *X*, precision($Y|X$)=$\frac{|X\cap Y|}{|Y|}$ and recall($Y|X$)=$\frac{|X\cap Y|}{|X|}$. Both are proportional to $|X\cap Y|$, the size of the set intersection of the pair of texts. The definition of the intersection is introduced by the aid of a grid, where every cell in the grid is the coordinate of some word in the candidate text with some word in the reference text. Whenever a cell in the grid coordinates two words that are identical is called a *hit*. Computing the match size as the number of hits in the grid would result in double counting. Therefore, the definition is based on the concept of "maximum matching" from graph theory [5]. A *matching* is a subset of the hits in the grid, such that no two hits are in the same row or column. The *match size* of a matching is the number of hits in the subset. A *maximum matching* is a matching of maximum possible size for a particular grid. The *maximum match size* (MMS) is the size of any maximum matching. The MMS is divided by the length of the candidate text (*C*) or the length of the reference text (*F*) to obtain the precision or the recall, respectively: precision($C|F$)=$\frac{MMS(C,F)}{|C|}$ and recall($C|F$)= $\frac{MMS(C,F)}{|F|}$.

In order to reward longer matches, a generalized definition of the match size is adopted; size(M)= $\sqrt[e]{\sum_{r\in M} length(r)^e}$, where *r* is a run, defined as a contiguous sequences of matching words appearing in the grid as diagonally adjacent hits running in parallel to the main diagonal. For *e*>1 computing MMS is NP-hard, therefore it is obtained using a greedy approximation that builds a matching by iteratively adding

the largest non-conflicting aligned blocks. The parameter *e* is adjusted to weighting matching longer runs. A typical value of *e* is 2. To account for multiple references, the references are concatenated in arbitrary order. Then the maximum matching is computed, with a barrier between adjacent references preventing runs to cross the barriers. Finally, the MMS is normalized with respect to the length of the input texts.

4.3 Adaptation of the Metrics to the Arabic Language

The described metrics have been primarily applied and customized for the English language. For instance, they provide the option to account for case sensitivity. While the Arabic language does not have case sensitivity, but it does have some other features that need to be accounted for. The evaluation metrics have been modified such that they can adapt to some peculiarities in the Arabic language, which are tolerated by human being. For instance, we consider the following cases:

- ا, أ, إ: It is quite common for people to write the letter ا, instead of أ or إ. Since this error is tolerated by human, we modified the evaluation metrics such that they take this into consideration as follows: if the candidate token includes an ا, while the corresponding token in the reference translation is with a hamza (أ or إ) for all references, the token is given a score α, where 0≤α≤1. If on the other hand, the candidate token includes a hamza (أ or إ) then it must match one reference token with the hamza in the same position, otherwise it is given a zero score.
- ى and ي: since mixing ى with ي is a common error that could be tolerated by humans, the modification entails giving a score 0≤α≤1 for a candidate token not matching a token in any reference because a ى mixed with ي or vice versa.
- ة and ه / ا and آ: mixing ة with ه or mixing ا with آ are considered errors that are not tolerated in the algorithm and are given a score of zero.

It should be noted that we do not account for all possible cases. Rather, we introduce the methodology that other special cases could follow to tune the metrics to suit the different levels of tolerance needed. The above cases are used only as examples implemented in our evaluation.

5 Datasets and Experimental Design

The experiments reported in this paper are conducted on datasets prepared from the EOLSS. Preparing our own test datasets stemmed from the desire to evaluate the UNL MT systems on real data sets and real applications. Further, there are no publically available datasets for the language pair English-Arabic as the ones available from NIST or Linguistic Data Consortium (LDC).

Experiments are conducted using data drawn from the EOLSS encyclopedia, which is used as the English corpus. The test dataset contains around 500 sentences, composed of 8220 words, drawn randomly from 25 documents containing around 15,000 sentences. The length of the test sentences varied; with a mean 16.44 and

standard deviation. The random selection ensured that the dataset covers the whole range of the 25 documents.

The output of the UNL system is evaluated and compared to other available systems supporting translation from English to Arabic. Three systems are considered: Google, Tarjim of Sakhr and Babylon.

Four reference translations have been prepared for the test dataset. Four professional translators were provided with the English sentences and they were requested to generate the Arabic translation, without being exposed to any output from MT system. The dataset was not split among different translators, that is each translator processed the entire dataset to ensure the same style within a reference.

Post edited versions of the UNL output have been prepared using human annotators and used in different experiments with dual purpose. One, is to evaluate the improvements introduced by post editing a machine output. Second, is to measure how far the UNL output is from the closest acceptable translation that is fluent and correct. A similar idea has been adopted in [21], where they create a *human-targeted reference*, a single reference targeted for this system, and shown to result in higher correlation with human judgment. Four post edited translations were prepared:

- PE-UNL: This form of post editing was performed by providing monolingual Arabic annotators with the Arabic output from the UNL MT and the UNL representation resulting from the encoding. The annotator was requested to correct the output by fixing any errors resulting from the lack of a grammar rule or the lack of semantic representation in a UNL expression.
- PE-En1 and PE-En2: This post editing was performed by providing bilingual annotators with the original English sentences, the UNL MT output and they were requested to perform the minimum changes needed to correct the output generated by the UNL such that the final output is fluent, grammatically correct and has the correct meaning relative to the English version. However, they were not requested to generate the best translation.
- PE-Pub: This post editing was conducted by an expert in the Arabic language, who was given the PE-En1 and was asked to render the sentence qualified for publishing quality, that is, making the article cohesive and ensuring that the sentence is written in a typical Arabic style.

Basic preprocessing was applied to all datasets. The sentences were tokenized, with removal of punctuation and diacritization. These preprocessing seemed essential as some systems use diacritization while others do not. Also some translators include different punctuation and diacritization, while others do not, or do with different degrees. Therefore, it seemed that removing them would result in a fairer comparison.

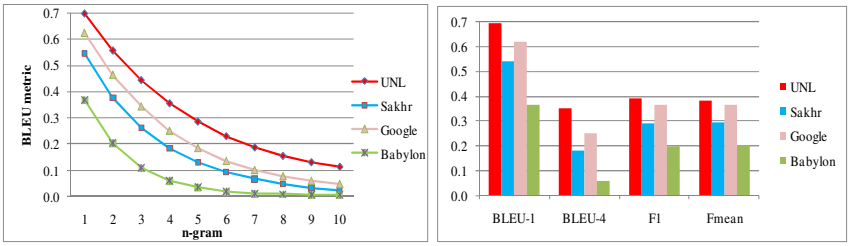
In the conducted experiments, the different parameters have been set as follows. For the BLEU, we use uniform weight w_n = 1/N and we vary the *n*-gram; although in some experiments we only show BLEU-1 and BLEU-4 reflecting adequacy and fluency respectively. For the F1 and F_{mean} metrics, the exponent *e* has been set to 2, which is the typical value used for weighting longer runs. For the adaptation introduced to the Arabic language, the parameter α has been set to 0.7.

6 Results

6.1. Evaluation using Professional Reference Translations

The dataset is evaluated for the three metrics using the four references obtained from the professional translators. One feature of BLEU not captured by unigram based metrics is the notion of word order and grammatical coherence achieved by the use of the higher level of *n*-grams. The *n*-grams are varied from 1 to 10 and BLEU has been computed for the four MT systems as shown in Figure 3(a). It is observed that UNL results in the best score, followed by Google, Sakhr, then Babylon. These results are statistically significant at 95% confidence. For n=1, which accounts for adequacy, it shows that all of the systems, except Babylon, provide reasonable adequacy, with UNL being the best. For higher *n*-grams, which captures fluency and the level of grammatical well formedness, as expected BLEU decreases as n increases. It is noted though that UNL provides better fluency than others. While on adequacy (n=1), UNL shows an improvement of 28% and 12% over Sakhr and Google respectively, for $4 \leq n \leq 10$, the improvement ranges from 42% to 144% over Google and 94% to 406% over Sakhr. For Babylon, it is observed that the decay is very fast, indicating the lack of fluency in its output.

When recall is taken into account, represented in F_1 and F_{mean}, Figure 3(b), it is noticed that UNL still outperforms all other, with significant improvement over Sakhr and Babylon, but with only marginal improvements over Google.



(a) **(b)**

Fig. 3. (a) BLEU metric for MT systems, varying *n*-grams (b) BLEU, F_1 and F_{mean} for MT systems. Results are obtained with professional human translation references.

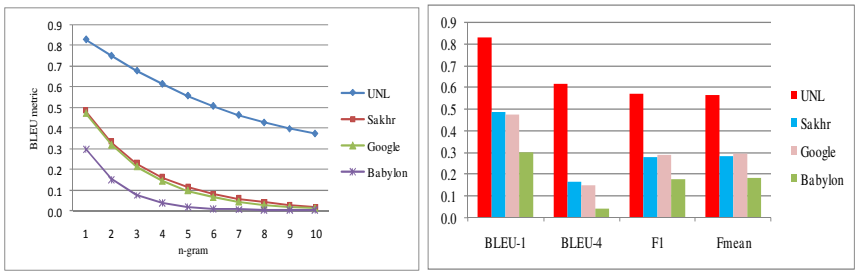
6.2. Evaluation using Post Edited References

In this experiment, we present the evaluation of the dataset making use of the post edited versions of the UNL output as references. They are sought to be good choices for references, since they are considered acceptable translations, yet, they are cheap to obtain. Also, they can be considered a possible substitute for subjective human judgment of MT quality. A similar approach has been adopted in [21].

From Figure 4, it is observed that the UNL is better than the three other systems; Google and Sakhr show similar performance while Babylon shows the poorest results. Although results are expected to be biased towards the UNL, it is observed that results follow the same trend as the ones obtained from the professional human translations.

Hence, the post edited versions could be considered a cheap and quick way of obtaining the tendency of the systems behavior.

It is worth mentioning that, analyzing the UNL performance with respect to its post edited versions gives an indication of how far it is from the closest acceptable translation. It is noted that the large values of BLEU, F_1 and F_{mean} for UNL is a good indicator that the output is not far off.



(a) **(b)**

Fig. 4. (a) BLEU metric for MT systems, varying *n*-grams (b) BLEU, F_1 and F_{mean} for MT systems. Results are obtained with post edited references.

6.3. Evaluating the Post Edited Translations as Systems

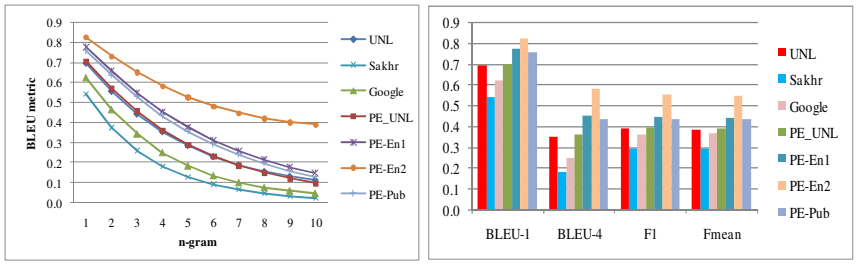
In this experiment, the four post edited versions of the UNL output are evaluated as systems output against the four professional human translation as references. This will give us an indication of how much improvements are obtained from post editing. Results are plotted in Figure 5 and show that all post editing versions result in improvements in all measures as compared to the raw output of the UNL, Google or Sakhr. In the following we analyze results against the UNL.

Considering PE-UNL, which is the cheapest, since it introduces only minor fixes comparing the UNL representation to the system output; it shows almost identical performance to the UNL output with improvements not exceeding 3% for all metrics.

Examining a more expensive post editing, namely PE-En1 and PE-En2, both of them yield an improvement. PE-En1 improves BLEU with a range from 12% to 41%, with higher improvements achieved for higher *n*-grams. Also it results in improvements around 15% for both F_1 and F_{mean}. PE-En2 on the other hand gives much higher improvements, ranging from 19% to 250% on BLEU and 42% on F_1 and F_{mean}. It should be noted that the qualifications of the two persons who performed the post editing were the same, so the degree of improvement obtained is subjective and needs to be weighed against its cost.

Turning to PE-Pub, which is the most expensive post editing, results are disappointingly low, especially in comparison to PE-En1 which was the source PE-Pub departed from. Since PE-Pub is a publishing quality; it ensures cohesion and typical Arabic style, which will result in removing structural interference such as cataphora, inappropriate nominal chunking or inappropriate coordination. For example, the English sentence "The management of freshwater resources" is translated by all systems and translators as "إدارة موارد المياه العذبة " which is a correct translation. However, the Arabic editor changed it to "موارد المياه العذبة وإدارتها" to

remove nominal chunks resulting from three successive nouns. This results in mismatch of PE-Pub with all references, hence, a low score. This implies that features such as cohesion and typical Arabic style are not captured by any of the MT metrics and more efforts needs to be exerted to devise metrics that account for these features.



(a) **(b)**

Fig. 5. (a) BLEU metric, varying *n*-grams (b) BLEU, F_1 and F_{mean}. Evaluating post edited versions as systems output

6.4. Responsiveness of the Systems to the Complexity of the Corpus

The test dataset has been categorized into three groups according to the difficulty of the sentences. Difficulty is judged by linguists based on the complexity of the structure of the sentence as well as its length. The first group G1 contains simple sentences, group G2 contains moderate sentences while group G3 contains complex sentences. The categorization by the linguists resulted in G1, G2 and G3 containing 50, 215 and 235 sentences respectively.

Figure 6, 7 and 8 plots BLEU, F_1 and F_{mean} for G1, G2 and G3 respectively. Results are shown along the values resulting from the global dataset. For space constraint we show BLEU while varying *n*-grams for G2 only in Figure 9. For G1, BLEU-3 is plotted and not BLEU-4, because BLEU-4 did not yield results as the length of the sentences was too short to produce 4-grams. It should be mentioned that the number of words in the Arabic language is less than the number of words in the English language as the Arabic language is an agglutinative language. Therefore, it is expected that sentences of group G1 would not be more than 3-word long.

Comparing the results of each group with its corresponding global value, it is noticed that for G1, the values are larger than the global value for all metrics, with the gap more noticeable for BLEU-3, F_1 and F_{mean}. For G2, the values are also larger than the global values for all metrics, with smaller differences than in the case of G1. However, for G3, the values are constantly lower than the global values. This implies that simple and moderate sentences yield high values for all metrics, while complex statements are the ones which results in low values.

Comparing the results of the 3 MT systems within the same group, it is observed that Google results in the best score for G1, with improvements reaching 42%. However, UNL shows higher values for G2 and G3 on all metrics reaching improvements of 23% and 52% over Google; 66% and 107% over Sakhr respectively. This implies that UNL outperforms Sakhr and Google in generation of sentences with complex structure.

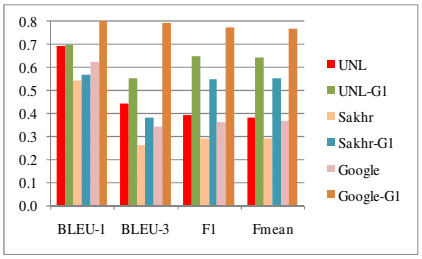


Fig. 6. BLEU, F_1 and F_{mean} for group G1

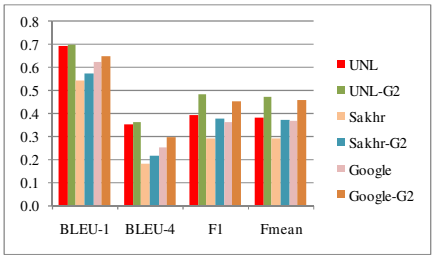


Fig. 7. BLEU, F_1 and F_{mean} for group G2

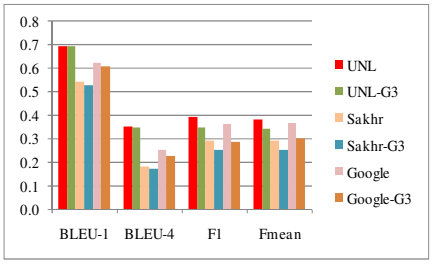


Fig. 8. BLEU, F_1 and F_{mean} for group G3

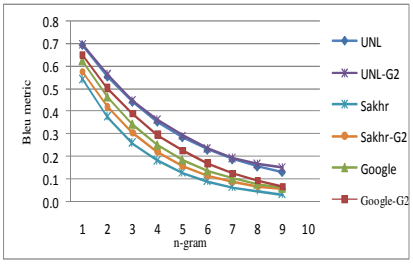


Fig. 9. BLEU metric, varying *n*-grams, for G2

7. Conclusions

In this research, we presented an evaluation for a MT system based on the UNL system. The evaluation has been conducted on the Encyclopedia of Life Support Systems (EOLSS). Three widely used automated metrics were evaluated, namely BLEU, F_1 and F_{mean}. The three metrics have been modified to adapt to some peculiarities in the Arabic language. The MT UNL system has been compared to other systems supporting English-Arabic translation, namely Google, Tarjim and Babylon. Results revealed that UNL performed better than the three systems on all metrics, especially when generating sentences with a complex structure. Evaluating annotated versions of the UNL output shown that they can be used as cheap references in order to highlight the tendency of the systems behavior. Results also revealed that current metrics do not capture features such as cohesion and typical Arabic style; hence, more work needs to be done in this direction. The framework of the evaluation presented will serve to analyze further development of the UNL MT system by comparing its output with suggested changes.

References

1. Agrawal, A., Lavie, A.: METEOR, M-BLEU and M-TER: Evaluation Metrics For High Correlation with Human Rankings of Machine Translation Output. In Proc. of the 3rd Workshop on Statistical Machine Translation, pp. 115-118, Ohio, June (2008)

2. Alansary, S., Nagi, M., Adly, N.: A Semantic Based Approach for Multilingual Translation of Massive Documents. In The 7th International Symposium on Natural Language Processing (SNLP), Pattaya, Thailand. (2007)
3. Alansary, S., Nagi, M., Adly, N.: Generating Arabic text: The Decoding Component of an Interlingual System for Man-Machine Communication in Natural Language. In the 6th International Conference on Language Engineering, 6-7 December, Cairo, Egypt. (2006)
4. Banerjee, S., Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proc of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization, Ann Arbor, (2005)
5. Cormen, T., Leiserson, C., Rivest R., Stein, C.: Introduction to Algorithms, 2nd Edition, MIT Press (2001)
6. Encyclopedia of Life Support Systems, http:://www.eolss.net/
7. Lavie, A., Agarwal, A.: METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Proceedings of the Second ACL Workshop on Statistical Machine Translation, pp 228–231, Prague, June. (2007)
8. Lavie, A., Pianesi, F., Levin, L.: The NESPOLE! System for Multilingual Speech Communication over the Internet. IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 5, Sept (2006)
9. Lavie, A., Sagae, K., Jayaraman, S.: The Significance of Recall in Automatic Metrics for MT Evaluation. In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), pp. 134–143, Washington, DC, Sept. (2004)
10. Lee, Y., Yi, W., Seneff, S., Weinstein, C.: Interlingua-based Broad-coverage Korean-to-English Translation in CCLINC. In Proceedings of the 1st International Conference on Human Language Technology Research, san Diego (2001)
11. Leusch, G., Ueffing, N., Ney, H.: CDER: Efficient MT Evaluation Using Block Movements. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. (2006)
12. Lin, C., Och, J.: ORANGE: a Method for Evaluation Automatic Evaluation Metrics for Machine Translation. COLING 2004, pp 501-507, Switzerland, Aug. (2004)
13. Lopez, A.: Statistical Machine Translation. In ACM Comp. Surveys, Vol. 40, Aug (2008)
14. Melamed, I., Green, R., Turian J.: Precision and Recall of Machine Translation. In Proceedings of the HLTNAACL 2003: pp. 61–63, Canada. (2003)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318, Philadelphia (2002)
16. Papineni, K., Roukos, S., Ward, T., Henderson, J., Reeder, F.: Corpus-based Comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In Proceedings of Human Language Technology 2002, San Diego, CA. (2002)
17. Przybocki, M., Sanders, G., Le, A.: Edit Distance: a Metric for Machine Translation Evaluation. In LREC (2006)
18. Rijsbergen, C. : Information Retrieval. Butterworths, London, 2nd edition, (1979)
19. Shaalan, K., Monem, A., Rafea, A., Baraka, H.: Mapping Interlingua Representations to Feature Structures of Arabic Sentences. The Challenge of Arabic for NLP/MT. International Conference at the British Computer Society, London,; pp.149-159. (2006)
20. Sinha, R. et al.: ANGLABHARTI : a Multilingual Machine Aided Translation from English to Indian Languages. IEEE Intl Conference on Systems, Man and Cybernetics (1995)
21. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul. J.: A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of AMTA, Boston, (2006)
22. Turian, J., Shen, L., Melamed, I.: Evaluation of Machine Translation and its Evaluation. In Proceedings of MT Summit IX. (2003)
23. Uchida, H., Zhu, M., Della Senta, T.: The Universal Networking Language. UNDL Foundation. (2005)