The GREYC/LLACAN Machine Translation Systems for the IWSLT 2010 Campaign

Julien Gosme[(1)], *Wigdan Mekki*[(1)], *Fathi Debili*[(2)], *Yves Lepage*[(1)], *Nadine Lucas*[(1)]

(1) GREYC, Université de Caen Basse-Normandie, Caen, France {firstname.lastname}@info.unicaen.fr

(2) LLACAN, UMR 8135, CNRS, Villejuif, France {firstname.lastname}@gmail.com

Abstract

In this paper we explore the contribution of the use of two Arabic morphological analyzers as preprocessing tools for statistical machine translation. Similar investigations have already been reported for morphologically rich languages like German, Turkish and Arabic. Here, we focus on the case of the Arabic language and mainly discuss the use of the G-LexAr analyzer. A preliminary experiment has been designed to choose the most promising translation system among the 3 G-LexAr-based systems, we concluded that the systems are equivalent. Nevertheless, we decided to use the lemmatized output of G-LexAr and use its translations as primary run for the BTEC_AE track. The results showed that G-LexAr outputs degrades translation compared to the basic SMT system trained on the un-analyzed corpus.

1. Introduction

We investigate the effect of morphological preprocessing on statistical machine translation quality. We focus on the case of the Arabic to English translation task, *i.e.*, the BTEC_AE track of IWSLT 2010 campaign. Arabic is a language with very rich morphology, the complexity of which challenges machine translation (see [1] for a description of some morphological problems relatively to this task). On the opposite, English morphology is quite poor. It thus makes sense to use the output of an Arabic morphological analyzer to possibly improve a statistical machine translation system.

Experiments in using morphological analysis to get improved translation quality have already been reported for German [2] or Turkish [3]. These works used various kinds of segmentation, lemmatization and POS tagging. In the case of Arabic, previous works led to opposite conclusions: on one hand, Lee [4] (affix-stem segmented Arabic) and Habash and Sadat [5] (linguistically informed tokenization) showed that morphological preprocessing may be helpful for statistical machine translation; on the other hand, Diab, Ghoneim and Habash [6] concluded that the use of morphological analysis led to no improvement (for partial vocalization) or even to worse results (with full vocalization).

In this paper, we similarly inspect the use of morphological analysis as a preprocessing step in Arabic-English statistical machine translation, but we use an in-house morphological analyzer, G-LexAr, and compare its performance with the well known Buckwalter's morphological analyzer (BAMA). As for translation results, our conclusion are similar to those in [6], as the contribution of morphological analysis is negative on the translation task: the performance degrades in comparison to a standard basic statistical machine translation system.

The paper is organized as follows: Section 2 succinctly presents the data and the preprocessing and postprocessing steps used in our setting. Section 3 describes the G-LexAr and Buckwalter's morphological analyzers and shows some of their differences. Section 4 presents the details of the design of a basic machine translation system and five different versions obtained from different outputs of the two previous analyzers. Section 5 presents preliminary results obtained with these different systems on newswire LDC data and the results obtained during this year evaluation campaign. Conclusions are presented in Section 6.

2. Preprocessing and postprocessing

2.1. Case and punctuation

The BTEC_AE data delivered by the organizers of the campaign consistently use natural writing standards in training set, development sets and test sets. Arabic is written as expected by native speakers: the sentences are segmented in hyperwords and they include non-tokenized punctuations. English texts are written accordingly including capitalized words and non-tokenized punctuations.

Our preprocessing step thus consisted in the tokenization of punctuation signs in Arabic and English texts by addition of spaces. As for English, in addition, we lowercased all texts.

Consequently, English outputs require a post-processing step which consists in re-introducing capitalization and removing extra spaces around punctuations. In this way, the English translations submitted for evaluation follow the same writing standards as the English data delivered by the organizers.

2.2. Encoding

Both analyzers we used, G-LexAr and Buckwalter's morphological analyzer, only handle Arabic encoded in Windows-

1256. For this reason, we had to convert the encoding of the Arabic texts released by the organizers from Unicode to Windows-1256 to be able to use the two previously mentioned morphological analyzers. The English text encoding was left untouched.

3. Morphological Analysis

In this section, we present the two morphological analyzers we used, with some more details for G-LexAr as this analyzer is not as popular as BAMA.

3.1. G-LexAr

G-LexAr is a program for the morpho-grammatical analysis of Arabic, classical and Modern Standard Arabic. Its output can include vocalization or not depending on the user's requirement. Its development started in 1982. The main ideas used in the development of G-LexAr are detailed in [7] (for instance, the comparison of Arabic POS tag sets). The analyzer takes one or several texts as input and it outputs their respective analysis, *i.e.*, words are segmented, vocalized, lemmatized, and labeled. It is based on the implementation of a large number of lexicons and rules. The emphasis is on processing speed improvement to the expense of storage space. It consists in three steps.

3.1.1. First step

The first step of the G-LexAr analysis segments the text into morphological units (simple forms and concatenated forms in Arabic script which are called hyper-forms) and filters other strings inputs that do not fall within the morphological analysis of Arabic itself.

3.1.2. Second step

The second step analyzes the hyper-forms independently of their contexts. For each hyper-form, it provides a tree with all of its possible segmentation, vocalization, lemmatization candidates, and all possible POS tags. Each morphological unit (MU) is assigned a lexical tree which can be represented as in Figure 1.

- In this figure, the lemmas are actually hyper-lemmas since they are associated with simple or concatenated forms. Similarly, the hyper-forms are linked with the "hyper-grammatical" categories since they are with basically concatenated forms.
- The main problem here is the agglutination of pronouns (enclitics), and articles, prepositions, conjunctions and other particles (proclitics). Multiple segmentations are possible and the problem is to identify the correct segmentation. To solve this problem, rules are applied that verify the compatibility of binary or ternary composition sequences of the form: proclitic + simple form + enclitic.

G-LexAr implements a lexicon of simple forms and several rule files. The lexicons contain 82,000 vocalized lemmas, which amounts to 1.5 million vocalized simple forms, and corresponding to 500,000 unvocalized simple forms for 66 proclitics and enclitics. In order to speed up the process, a new architecture has recently been developed. The lexicon of hyper-forms is implemented in such a way that each entry is directly assigned its corresponding lexical tree.

3.1.3. Third step

The third step performs the grammatical tagging and pruning of the lexical trees obtained as output of the second step. The tagging consists in providing the entire list of ambiguities (all potential tags of a word), rather than providing the two or three commonly allowed tags. Pruning comes next: it retains only those branches which leaves correspond to the selected tags. To make this process faster, pruning implements a lexicon that directly associates pruned lexical trees to pairs of words and tags. Such an architecture leads to a ten-times faster processing measured on the same machine for the same linguistic performance, in comparison to an older implementation where pruning was actually computed during analysis.

The final output of G-LexAr is provided in XML[1] format as shown in the following example:

```
<mot UM="بسرعة"
     Langue="A" Taille="5" Position="142" Etat="10">
  <AMG V="بِسُرْعَةٍ"
       L="بِ+سُرْعَة ـ ـ ـ ـ ـْ ـَ ـْ+"
       HCG="إسم منوّن مجرور+حرف جرّ"/>
  <AMG V="بِسُرْعَةِ"
       L="بِ+سُرْعَة ـ ـ ـ ـ ـْ ـَ ـْ+"
       HCG="مضاف مجرور+حرف جرّ"/>
  <AMG V="بِسَرِعَةِ"
       L="بِ+سَرِع ـ ـ ـ * ـْ ـَ ـْ+"
       HCG="نعت مضاف مجرور+حرف جرّ"/>
                  ⋮
</mot>
```

For each word (XML element `mot`), G-LexAr gives several possible grammatical analyses (XML element `AMG`) sorted by decreasing ranking of relevance. Each analysis gives:

- a vocalization of the word (attribute `V`) ;
- a lemmatization (attribute `L`) with its segmentation following the representation given as: proclitic + base + enclitic
- a POS tagging (XML element `HCG`), including a list of grammatical categories (such as noun, verb, etc.) separated by the plus sign (+).

[1] eXtensible Markup Language.

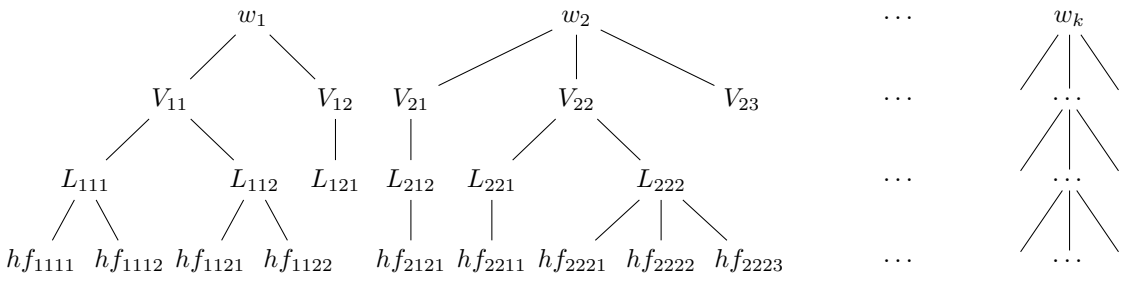

Figure 1: G-LexAr second step output. Words, vocalized forms, lemmatized forms and hyper-forms are written w_i, V_{ij}, L_{ijk} and hf_{ijkl} respectively.

3.2. Buckwalter Morphological Analyzer

The Buckwalter Morphological Analyzer (BAMA) uses a concatenative lexicon-driven approach where morphotactics and orthographic rules are built directly into the lexicon itself instead of being specified in terms of general rules that interact to realize the output [8]. For each input string, the analyzer provides a solution (in Buckwalter Transliteration), including the words unique identifier or lemma ID, a breakdown of the constituent morphemes (prefixes, stem, and suffixes), and their POS values and corresponding English glosses, as in the following example:

```
INPUT STRING: العاز
 SOLUTION 1:  >alogAz
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN
       GLOSS: mysteries/enigmas
 SOLUTION 2:  >alogAzu
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+u/CASE_DEF_NOM
       GLOSS: mysteries/enigmas + [def.nom.]
 SOLUTION 3:  >alogAza
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+a/CASE_DEF_ACC
       GLOSS: mysteries/enigmas + [def.acc.]
 SOLUTION 4:  >alogAzi
    LEMMA_ID: lugoz_1
         POS: >alogAz/NOUN+a/CASE_DEF_GEN
       GLOSS: mysteries/enigmas + [def.gen.]
                  ⋮
```

4. Translation Systems Compared

We designed 6 different translation systems for the BTEC_AE track. All translations are performed by the phrase-based statistical machine translation toolkit Moses [9] which makes use of the word aligner GIZA++ [10]. The different translation systems use different forms of Arabic input texts, which correspond to the different possible outputs of the G-LexAr and BAMA morphological analyzers. We also designed a basic translation system which makes use of untouched un-analyzed Arabic text.

The G-LexAr analyzer output includes vocalized, lemmatized and segmented Arabic formats, while BAMA output includes lemmatized and segmented forms only. Both analyzers output a list of analysis for each token (hyper-form) ordered by relevance scores. In this work, we always choose the best solution for each token as the unique analyzed form.

As for G-LexAr, the vocalized format corresponds to standard Arabic texts that would have been entirely vocalized, as is the case for the Qur'an or children books. The segmented format is a text where all hyperforms have been split into simple forms with their surrounding proclitics and enclitics separated by spaces. The lemmatized format differs from the segmented format in that simple forms are lemmatized, *i.e.*, replaced by their corresponding lemmas; in addition, proclitics and enclitics are removed. As an example, Sentence 3 in Table 2 shows that articles are removed in the lemmatized format while they are kept in the segmented format.

The 6 Arabic types of input texts that are obtained without or with morphological analysis can be summarized in the following table:

A Analyzer	F Format
none	original
G-LexAr	vocalized
	lemmatized
	segmented
BAMA	lemmatized
	segmented

Translation systems are built according to the following steps:

- Analyze Arabic parts of training set, development sets 1, 2, 3, 6 and 7 and 2009 and 2010 test sets with analyzer A and retain the best F form for each token. Leave English parts of training set and development sets untouched.
- Train a machine translation (Moses and GIZA++) by using the training set prepared in step 1.
- Tune the machine translation with all the development sets prepared in step 1.
- Translate tests sets prepared in step 1 with Moses' decoder.

This protocol allows us to compare the benefit of several Arabic morphological analysis forms and strategies.

Table 1 shows statistics for the morphological analysis outputs of the training corpus. These figures show that G-LexAr and BAMA follow different morphological analysis strategies which for instance lead to more or less segmentation. As examples, Table 2 gives the outputs of the analysis of three different sentences performed by G-LexAr and BAMA according to the previous five different schemes given in the previous table.

Table 1: Statistics on the 6 different types of Arabic texts (training corpus). See Table 2 for examples of the corresponding formats.

Arabic format		Size (Mb)	# tokens	tokens per line
untouched (original)		1	159,006	8.0
G-LexAr	vocalized	2	159,006	8.0
	lemmatized	1	159,006	8.0
	segmented	1	203,338	10.2
BAMA	lemmatized	1	159,000	8.0
	segmented	1	255,948	12.8

Table 2: Examples of G-LexAr and BAMA analysis. This table shows 3 Arabic sentences analyzed (sentence 3, 4 and 5 of the 2010 test set). As for Arabic speakers, reading the outputs of the G-LexAr analyzer is of course much easier than reading the Buckwalter's transliteration.

no morphological analysis		
original	3	ما هي تكلفة السباحة ليوم كامل في حمام السباحة ؟
	4	أود المبيت هنا الليلة . هل لديكم فراش ؟
	5	جيد . أرجو تعبئة هذه الاستمارة .
G-LexAr		
vocalized	3	مَا هِيَ تَكْلُفَةِ السَّبَّاحَةِ لِيَوْمِ كَامِلِ فِي حَمَامِ السَّبَّاحَةِ ؟
	4	أَوَدَّ الْمَبِيتُ هُنَا اللَّيْلَةُ . هَلْ لَدَيْكُمْ فَرَاشٌ ؟
	5	جَيِّدٍ . أَرْجُوَ تَعْبِئَةِ هذه الْإِسْتِمَارَةِ .
lemmatized	3	مَا هِي تَكْلُفَة سَبَّاحَة يَوْم كَامِل فِي حَمَام سَبَّاحَة ؟
	4	أَوَدّ مَبِيت هُنَا لَيْلَة . هَلْ لَدَى فَرَاش ؟
	5	جَيِّد . رَجَا تَعْبِئَة هذه إِسْتِمَارَة .
segmented	3	مَا هِي تَكْلُفَة الْ سَبَّاحَة لِ يَوْم كَامِل فِي حَمَام الْ سَبَّاحَة ؟
	4	أَوَدّ الْ مَبِيت هُنَا الْ لَيْلَة . هَلْ لَدَى كُمْ فَرَاش ؟
	5	جَيِّد . رَجَا تَعْبِئَة هذه الْ إِسْتِمَارَة .
BAMA		
lemmatized	3	mA hiya takolif sab¯AH yawom kAmil fiy HamAm sab¯AH ?
	4	>awid mabiyt hunA layol . hal ladayo firA$?
	5	jay¯id . rojuw taEobi} h'*ihi {isotimAr .
segmented	3	mA hiya takolif ap Al sab AH ap li yawom kAmil fiy HamAm Al sab AH ap ?
	4	>awid a Al mabiyt hunA Al layol ap . hal ladayo hi kum firA$?
	5	jay¯id . >a rojuw taEobi} ap h'*ihi Al {isotimAr ap .

5. Results

5.1. Preliminary experiment

The goal of our participation to the evaluation campaign was to evaluate the possible contribution of the G-LexAr analyzer to statistical machine translation using the popular off-the-shelf tools MGIZA++/Moses. We thus conducted a preliminary experiment in order to choose the most promising G-LexAr-based system as our primary system for the IWSLT Arabic-English task. The protocol used is described in section 4. The data we used for that was a sample of 251,000 Arabic-English parallel sentences corpus released by the LDC (newswire).[2] Table 3 shows the scores obtained.

Since all scores seem very close, we conducted statistical significance tests using all metrics except NIST.[3] We applied the method based on bootstrap re-sampling described in [11]. For each test corpus of 500 sentences, we constructed 1,000 new test corpora of 500 sentences each by uniform sampling with replacement. Using this method, we obtained 95% confidence interval scores (assuming 1,000 tests corpora is large enough). The results are shown in Table 4.

For each given metric, all interval scores for G-LexAr, BAMA and original overlap. This is indicated by the non empty intersections in the table. This shows that the small differences observed in Table 3 are not statistically significant, a conclusion which cannot be drawn by just comparing median scores. Therefore, we concluded that we can choose any G-LexAr Arabic morphological format as primary system for the IWSLT Arabic-English task of this year's evaluation campaign as it is not significantly different to the other systems trained on other formats. We choose the G-LexAr lemmatized format for our primary run. Table 5 summarizes the formats used for submitted run.

Table 3: Comparison of G-LexAr, BAMA and original (no Arabic morphological analysis performed) in the Arabic to English translation task. — Systems trained on a sample of 251,000 Arabic-English parallel sentences released by the LDC.

		mWER	NIST	BLEU	TER
original		0.4876	5.9833	0.2121	0.8244
G-LexAr	vocalized	0.4962	5.7590	0.1979	0.8399
	lemmatized	0.4999	5.7289	0.1975	0.8449
	segmented	0.4824	5.9623	0.2064	0.8169
BAMA	lemmatized	0.4868	5.9930	0.2092	0.8112
	segmented	0.4821	5.7312	0.1957	0.8431

5.2. IWSLT 2010 evaluation campaign

The results obtained on each of the different systems are given in Table 6. The best scores are in boldface characters for each metric. These results show that the translation

[2] ISI Arabic-English Automatically Extracted Parallel Text, LDC Catalog No. LDC2007T08

[3] The method used for significance test purposes can't be applied on the NIST metric because NIST does not give individual score for each sentence.

Table 4: Scores obtained by the translation systems based on G-LexAr and BAMA analysis and the untouched Arabic texts. Scores are presented as median scores and as 95% confidence interval scores (median [lower limit, upper limit]).

		mWER	BLEU	TER
original		0.4874 [0.4772, 0.4985]	0.2121 [0.1990, 0.2250]	0.8239 [0.8032, 0.8480]
G-LexAr	vocalized	0.4962 [0.4855, 0.5071]	0.1978 [0.1847, 0.2113]	0.8394 [0.8175, 0.8634]
	lemmatized	0.5000 [0.4896, 0.5106]	0.1973 [0.1850, 0.2092]	0.8451 [0.8237, 0.8699]
	segmented	0.4823 [0.4722, 0.4929]	0.2066 [0.1850, 0.2092]	0.8165 [0.7955, 0.8400]
BAMA	lemmatized	0.4869 [0.4774, 0.4972]	0.2091 [0.1963, 0.2214]	0.8111 [0.7905, 0.8332]
	segmented	0.4822 [0.4721, 0.4924]	0.1957 [0.1835, 0.2091]	0.8430 [0.8208, 0.8689]
	intersection	[0.4896, 0.4924]	[0.1990, 0.2091]	[0.8237, 0.8332]

Table 6: GREYC machine translation scores obtained in the BTEC_AE task.

morph. analyzer	form (run)	case/punc	BLEU	METEOR	f1	Prec.	Recl.	WER	PER	TER	GTM	NIST
none	original	yes/yes	**0.408**	0.693	**0.742**	**0.776**	0.711	**0.414**	0.371	**35.67**	0.703	6.855
	(contrastive1)	no/no	0.376	0.646	0.697	0.737	0.662	0.477	0.421	40.32	0.658	6.566
G-LexAr	vocalized	yes/yes	0.206	0.488	0.604	0.733	0.514	0.573	0.537	47.64	0.555	2.433
	(contrastive2)	no/no	0.173	0.413	0.525	0.674	0.430	0.669	0.618	53.42	0.484	1.598
	lemmatized	yes/yes	0.296	0.591	0.669	0.726	0.620	0.516	0.464	43.57	0.611	4.987
	(primary)	no/no	0.258	0.531	0.612	0.679	0.557	0.593	0.523	49.07	0.557	4.441
	segmented	yes/yes	0.287	0.607	0.673	0.706	0.643	0.525	0.460	44.76	0.628	5.555
	(contrastive3)	no/no	0.246	0.548	0.617	0.656	0.583	0.598	0.516	50.31	0.580	5.246
BAMA	lemmatized	yes/yes	0.391	0.700	0.738	0.749	0.728	0.428	**0.370**	36.52	0.692	7.157
	(contrastive5)	no/no	0.455	0.693	0.731	0.755	0.708	0.435	0.374	35.83	0.710	**7.736**
	(segmented)	yes/yes	0.386	**0.717**	0.736	0.717	**0.757**	0.438	0.376	38.50	**0.716**	7.434
	(contrastive4)	no/no	0.352	0.672	0.692	0.673	0.713	0.505	0.423	43.88	0.678	7.544

Table 5: Arabic formats used for the submitted runs.

Run	Morph. analyzer	Format
primary	G-LexAr	lemmatized
contrastive1	none	original
contrastive2	G-LexAr	vocalized
contrastive3	G-LexAr	segmented
contrastive4	BAMA	segmented
contrastive5	BAMA	lemmatized

systems on an untouched corpus outperforms all G-LexAr-based systems. In addition, the translation systems based on BAMA also outperform G-LexAr based systems.

The results of these experiments can be summarized as follows. The translation systems based on different G-LexAr output formats do not improve translation quality on newswire texts (preliminary experiment) and BTEC texts. This was also the case for BAMA, in the case/punctuation settings (natural texts), but not without punctuation and case-insensitive setting for evaluation, where the scores obtained are the best ones in BLEU over all formats.

The scores of the systems using G-LexAr formats may be explained when considering that the G-LexAr lexicons have been extracted from classical dictionaries. Such dictionaries do not contain modern words used in daily life such as: fax, taxi, hamburger, etc. On one hand, because of its domain, the tourism domain, BTEC is known to contain plenty of such daily life words. On the other hand, the LDC corpus, which consists in newswire texts, contains much fewer words of this type. In order to sustain this hypothesis, we checked the list of unknown words output by G-LexAr during the analysis of the BTEC training corpus. From this list, we selected the following 10 frequent lemmas that are modern usage words:

Arabic modern word	English meaning
أتوبيس	*autobus*
تاكسي	*taxi*
ويسكي	*whisky*
مترو	*metro*
ديسكو	*disco*
فلاش	*flash*
بيسبول	*baseball*
فاكس	*fax*
بيتزا	*pizza*
هامبورجر	*hamburger*

We found 35 unknown hyperforms derived from those lemmas, with a total number of occurrences of 238 in the training set.

That is to say, each such word appears in average 6.8 times. In comparison to the whole set of unknown words where each word appears 2.3 times in average (1,544 words

Table 7: Examples of translations: the first 10 sentences of the 2010 test set segmented by G-LexAr and BAMA along with English translation outputs using these formats.
Both analyzers allow the system to output the same translation candidate for sentence 1 and 3.
In sentence 2 and 4, the system using BAMA gives a more precise translation. However, both systems fail to translate *spend the night*.
While the system using G-LexAr fails to translate *please, fill the form* in 5,the system using BAMA provides *I hope packing this form*; but neither system succeed to provide a meaningful translation.
Even though, both systems using G-LexAr and BAMA do not provide the best translation for *spend the night* in 7 and 10, BAMA still provides a better translation. In sentence 10, the system using G-LexAr translates *Im* where the system using BAMA translates *we*.
In total, this table shows that the system using BAMA provides overall better translations than the system using G-LexAr.

#	G-LexAr/segmented (contrastive3)	BAMA/segmented (contrastive4)
1	ساعد ني . help me .	sAEad a niy . help me .
2	أصاب بال برد . had cold .	>aSab tu bi Al barod . i' ve got a cold .
3	ما هي تكلفة ال سباحة ل يوم كامل في حمام ال سباحة ؟ how much is the swimming for a full day at the swimming pool ?	mA hiya takolif ap Al sab¯AH ap li yawom kAmil fiy HamAm Al sab¯AH ap ? how much is it for a full day swim in the swimming pool ?
4	أود ال مبيت هنا ال ليلة . هل لدى كم فراش ؟ the مبيت أود here tonight . do you have bed ?	>awid a Al mabiyt hunA Al layol ap . hal ladayo hi kum firA$1 ? i' d like to mabiyt here tonight . do you have bed ?
5	جيد . رجا تعبئة هذه ال استمارة . good رجا . packing هذه form .	jay¯id . >a rojuw taEobi$1 ap h'*ihi Al $1isotimAr ap . all right . i hope packing this form .
6	رجي ال حاضر للبروفات بعد ثلاثة شهر و ستة شهر من الآن . س كان ال بآذل جاهزة خلال أحد عشر شهر . come رجي للبروفات and six months after three months from now . i' ll be ready in a suit ten a month .	ya rojaY Al HuDuwr li Al bruwf At baEoda valAv ap >a$1ohar a wa sit¯ ap >a$1ohar a min Al|na . sa ta kuwn Al ba*ol ap jAhiz ap xilAla >aH¯ad a Ea$1¯ar a $1ahar A . i come to the rojay bruwf are after three months and six months from now . you' ll be ready in the a suit eleven a .
7	أين أنا ؟ where we ?	>ayona naHonu ? where are we ?
8	أسف ل عدم جريء ال مكالمة على ال نحو الذي اتصل ب ه . رجا ال تحقق من ال رقم و معاودة ال اتصال أو ال اتصال ب مشغل ال شبكة ال محلي . we' re sorry for the call on the جريء not so are you calling رجا . the number and check معاودة call or an شبكة call the local .	na >osaf li Eadam >ujarA' Al mukAlam ap EalaY Al naHow Al¯a*iy $1it¯aSal tu bi hi . na rojuw Al taHaq¯uq min Al raqom wa muEAwad ap Al $1it¯iSAl >aw Al $1it¯iSAl bi ma$1ogal Al $1abak ap Al maHal¯iy¯ ap . we' re sorry for the call on eadam >ujara' like do you have . we have , please check the number and mueawad or a call to call the operator network local .
9	هل أمكن ك اعطائي تذكرة ال طيران خاصة ك ؟ do you possible اعطائي special my airline ticket ?	hal yu makonik ¡iEoTA$1 iy ta*okir ap Al TayarAn xAS¯ ap ka ? can you give me a flight for you ?
10	أنا هنا لل مبيت . we' re here for مبيت .	>anA hunA li Al mabiyt . i' m here to the mabiyt .

used 3,579 times), modern unknown words thus appear 3 times more. This points out the inadequacy of using G-LexAr to handle the BTEC texts. This problem was not revealed on the LDC corpus during our first translation experiments. A manual inspection of the morpho-syntactical analyses delivered by G-LexAr did not allow us to point at any major error of the analyzer. The result of this manual inspection is that the lowest scores obtained by the system using G-LexAr are mainly due to an out-of-vocabulary (OOV) problem, a problem that should be remedied by the indexation of such modern daily life terms into the analyzer lexicons.

As examples of outputs, ten translation candidates obtained with G-LexAr and BAMA segmentations are given and commented in Table 7.

6. Conclusion

Our investigation in using Arabic morphological analyzers for statistical machine translations mainly confirms the work of Diab, Ghoneim and Habash [6]. It allows us to conclude that Arabic morphological analyzers would not be very helpful to improve statistical machine translation systems in general. In this respect, BAMA and G-LexAr behave differently since the former does not really degrade translation system performance but the latter clearly does on BTEC texts.

From our manual inspection of results, we concluded that the clear differences observed in the results obtained on the BTEC texts used for the evaluation campaign (short sentences, small corpus) and on newswire texts from the LDC (longer sentences, larger corpus) come from insufficient coverage of the vocabulary. We spotted that out analyzer, G-LexAr, lacks daily life terms and was thus not able to handle the BTEC texts as efficiently as the newswire texts from the LDC. Out-of-vocabulary words appeared to constitute a real bottleneck for the use and contribution of this Arabic morphological analyzer.

Another result is that, surprisingly enough, the worst scores over all different G-LexAr formats were obtained for the vocalized format, which, expectedly, would have corresponded to some disambiguated form of the texts. Such a phenomenon has already been observed by [6] but was said to be possibly predictable since vocalization is far from being a solved issue yet.

For the most natural setting (case-sensitive and right punctuation) of evaluation, a disappointing conclusion is that the best scores were obtained with a system based on untouched texts. With the outputs of this translation system as our primary submission, we would have been ranked 7th instead of 11th out of 12 participants.

As a general conclusion to the experiments reported here, the question whether human-oriented vocalization, segmentation or lemmatization is really helpful for statistical machine translation, is still pending.

7. References

[1] N. Habash, *Arabic Computational Morphology*, ser. Text, Speech and Language Technology, 2007, vol. 38, ch. Arabic Morphological Representations for Machine Translation, pp. 263–285.

[2] S. Nießen and H. Ney, "Statistical machine translation with scarce resources using morpho-syntactic information," *Computational linguistics*, vol. 30, no. 2, pp. 181–204, 2004.

[3] A. Bisazza and M. Federico, "Morphological preprocessing for Turkish to English statistical machine translation," in *Proceedings of the International Workshop on Spoken Language Translation*, 2009, pp. 129–135.

[4] Y. Lee, "Morphological analysis for statistical machine translation," in *Proceedings of HLT-NAACL 2004: Short Papers on XX*. Association for Computational Linguistics, 2004, pp. 57–60.

[5] N. Habash and F. Sadat, "Arabic preprocessing schemes for statistical machine translation," in *Proceedings of the Human Language Technology Conference of the NAACL*. Association for Computational Linguistics, 2006, pp. 49–52.

[6] M. Diab, M. Ghoneim, and N. Habash, "Arabic diacritization in the context of statistical machine translation," in *Proceedings of MT-Summit*, 2007.

[7] F. Debili and E. Achour, H. et Souissi, "De l'étiquetage grammatical à la voyellation automatique de l'arabe," *Correspondances*, vol. 71, pp. 10–28, 2002.

[8] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0. LDC catalog number LDC2002L49," ISBN 1-58563-257-0, Tech. Rep., 2002.

[9] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, june 2007.

[10] F. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[11] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP*, vol. 4, 2004, pp. 388–395.