Evaluation Methodology and Results for English-to-Arabic MT

Olivier Hamon **Khalid Choukri**
ELDA
55-57 rue Brillat-Savarin
75013 Paris, France
{hamon;choukri}@elda.org

Abstract

This paper describes the evaluation campaign of the MEDAR project for English-to-Arabic (EnAr) MT systems. The campaign aimed at establishing some basic facts about the state of the art for MT on EnAr, collecting enough data to better train and tune systems and assessing the improvements made. The paper details the data used and their formats, the evaluation methodology and the results obtained by the systems. We conclude by giving some recommendations on MT evaluation for EnAr direction in terms of technology and resources.

1 Introduction and Objectives

When working with Arabic, most of the evaluation campaigns or Machine Translation (MT) systems only consider the Arabic-to-English direction. One of the major goals of MEDAR [1] has been to develop research around the English-to-Arabic direction, targeting several objectives:

- Develop a framework for the evaluation of English-to-Arabic MT systems;
- Develop baseline systems with background from existing open source tools;
- Produce data for MT training;
- Produce data for MT evaluation;
- Evaluate MEDAR baseline MT systems and compare them with other MT systems;
- Create and federate a new community around the English-to-Arabic MT theme;
- Make available a package containing the full set of resources from MEDAR.

In this paper, we describe the evaluation methodology and results of the MEDAR evaluation campaign for English-to-Arabic MT systems. Our very first goal is to identify the performance level of the MEDAR baseline systems developed within the project[2].

The evaluation is conducted in three phases. The first phase aims at setting some basic facts about the state of the art for MT on English-to-Arabic, including the development of baseline systems. The second phase aims at collecting enough data to better train and tune the systems. The third phase is the assessment of systems and their improvement.

In the following sections, we first present the baseline MT systems developed within the project, then the data used and their production design. Following that, we present the preparation of the evaluation campaign and the results of the systems. Finally, we draw some conclusions and describe the lessons learnt during the project. We also give some recommendations on English-to-Arabic MT evaluation in terms of technologies and resources.

Our evaluation has been setup in two phases. First, a dry-run has been carried out so as to test and check the evaluation protocol, then the effective evaluation of the MT systems has been realized.

2 MEDAR Baseline MT Systems

In MEDAR, two baseline Statistical Machine Translation (SMT) systems have been used. They have been developed by the University of Balamand ("Baseline1" hereafter) and IBM Egypt with the help of DCU ("Baseline 2" hereafter), respectively, and based on the open-source SMT

[1] http://www.nemlar.org

[2] These systems are available from the project website.

system Moses [3] (Koehn et al., 2007). The two baseline systems have been adapted to English-to-Arabic and Arabic-to-English directions. Moses has been chosen among other open-source MT systems because of its success on different languages. Details on the systems may be found in (Hamon & Choukri, 2010).

2.1 MEDAR Baseline 1

First, a baseline system has been built using Moses on English-to-Arabic and Arabic-to-English directions, and then it has been improved by setting up different parameters of the system. Several specificities of the Arabic language have been taken into account, starting with the preprocessing step where the tokenization, punctuation and lack of uppercase are handled. The different Moses tools have been used and adapted to Arabic, such as SRILM [4] (Stockle, 2002) or GIZA++[5] (Och & Ney, 2003). The University of Balamand has also developed a limited morphological analyser (Ghaoui et al., 2005) that separates prefixes and suffixes from the words and considers them as independent words. After translation, prefixes and suffixes are rearranged with their corresponding words. Finally, since synonyms were identified as a major source of errors, different suffixes have been added in the training to the English words, depending on the translated synonyms. During the decoding, words with equivalent synonyms in the target language are appended with the different synonyms and the phrase translation with the highest score is kept.

2.2 MEDAR Baseline 2

The second baseline system is also based on Moses. It uses SRILM and GIZA++ as translation model for word alignments and heuristics to build phrase table. The decoder used is the stack decoding algorithm. Preprocessing is made using a) the ArabicSVMTools package, which takes a regular transliterated Arabic text file and provides it tokenized, PoS-tagged and base phrase chunked, and b) the OpenNLP toolkit that performs sentence detection, tokenization, PoS-tagging, chunking, parsing, named-entity detection, etc.

[3] http://www.statmt.org/moses/
[4] http://www-speech.sri.com/projects/srilm/
[5] http://code.google.com/p/giza-pp/

3 Material

Each corpus is encoded in XML and UTF-8 and segmented into sentences. Moreover, MEDAR defined guidelines so as to produce high-quality reference translations and validate them.

3.1 Monolingual Training Data

Three sources have been used to produce the MEDAR monolingual corpus (see Table 1). Part of the corpus is provided by LDC[6], who has kindly shared some of its data for the purpose of evaluation only. Other data come from ELRA[7] or have been developed within MEDAR. Data from catalogues have just been converted to comply with our format as they were already clean without any garbage. Finally, 6 monolingual corpora have been produced within the project.

Name	Size [words]
Islamonline	20M
Wikipedia	31M
Wikibooks	1M
Wikinews	129M
Wikiquote	144M
Wikisource	69M
(ELRA) An-Nahar	113M
(ELRA) Al-Hayat	38M
(ELRA) LMD	475K
(ELRA) NEMLAR	494K
(LDC) Arabic Gigaword 4th Ed.	2GB

Table 1. Monolingual data used for training.

The IslamOnline corpus is made up of newspaper articles which have been crawled, cleaned and formatted according to the MEDAR requirements. Wiki raw data have been downloaded from the "Database Dump" of Wikipedia [8] then formatted following the MEDAR format without any further cleaning, the data being provided without garbage content. Most of the data are available either for R&D or for MEDAR evaluation purposes.

[6] http://www.ldc.upenn.edu
[7] http://catalogue.elra.info
[8] http://download.wikipedia.org/backup-index.html

3.2 Parallel Training Data

Three sources have been used to produce the MEDAR parallel corpus (see Table 2). LDC provided parallel data from its catalogue. The format of this data remains unchanged as it is compliant with the MEDAR requirements. A MEDAR corpus was built using the dry-run corpus. It consisted of a set of documents together with the 4 reference translations formatted into 4 parallel corpora of 10K words. Finally, two parallel corpora have been selected from already existing data: Meedan translation memory [9], composed of news articles, and the UN corpus[10] composed of collections from the United Nations General Assembly Resolutions. Crawling, cleaning and formatting have been made using our own scripts since the task was quite simple.

Name	Size [words]
MEDAR Dry-run	10K
Meedan	426K
UN	2,7M
(LDC) Multiple-Trans. Ar. Part 1	23K
(LDC) Ar. News Trans. Text Part 1	441K
(LDC) Multiple-Trans. Ar. Part 2	15K

Table 2. Parallel data used for training.

3.3 Evaluation data

To proceed with the testing of the systems, both a test corpus and a "masking" corpus must be built. The former allows scoring the systems against reference translations which are made by humans. The latter is much larger, not parallel, and is used to keep unknown the test corpus to the participants. Once the participant submissions have been received, only the part corresponding to the test corpus is kept. Input data are English texts coming from the Climate Change domain. The overall evaluation data has been built as follows:

1. Evaluation data is collected from many websites focusing on Climate Change;
2. Part of this test data, a test corpus, is selected to evaluate the MT systems;
3. The remaining words are used as a masking corpus;
4. The test corpus is translated four times by four different translators following translation guidelines, then validated following validation guidelines.

For the dry-run, during which the protocol is tested, the evaluation data are composed of about 210K running words: 10K words are used as test corpus, the rest being the "masking" corpus.
For the evaluation campaign, the evaluation data consist of about 40K words, where 10K are used as test corpus and the other 30K words as masking corpus. We decided to reduce the masking corpus size after the dry-run experience since participants had a short delay to produce the translation and the evaluation data was already large enough.

3.4 Scoring Tools and Methodology

Systems are evaluated using both automatic and human evaluations. Automatic scoring is done with BLEU (Papineni et al., 2001). Human evaluation is done using an interface developed at ELDA. For all the systems, each sentence is evaluated in relation to adequacy and fluency measures. For the former, the target sentence is compared to a reference sentence. For the latter, only the syntactic quality of the translation is judged. Judges grade all the sentences on a five-point scale where only extreme cases are explicitly defined, firstly according to fluency and then according to adequacy: both measures are done independently.
Two evaluations are carried out per sentence and done by two different judges, and sentences are distributed to judges randomly.

4 Dry-run

As a dry-run, there was no training of systems. None of the MEDAR MT systems receive any particular training either and a very basic data set has been used. Therefore, participants were free to use any kind of data they could obtain. As a consequence, systems are not directly comparable.
The MEDAR dry-run has been carried out in January and February 2010. Participants had 10 days to produce the output of their systems. Automatic and human results have been sent back to participants 5 and 10 days later, respectively.

[9] http://news.meedan.net/
[10] http://www.uncorpora.org

4.1.1 Participating Systems

Both MEDAR MT systems have been used. Since the evaluation campaign was open to external participants, so was the dry-run. In order to get participants, the campaign has been announced through several channels: mailing lists, personal contacts, networking, conferences, etc. Four participants replied and five submissions were made. The lack of participation may be due to the short delay between the start of the dry-run and the scoring. For comparison purposes, we used two online systems: Google Translate[11] and Systranet[12].

4.1.2 Results

Automatic results have been computed using four references. To compare what a human translator can do and put into perspective the results of the MT systems, one reference translation has also been compared to the other three.

For the human evaluation, 12 submissions have been evaluated: those from the 10 systems, plus 2 systems for which remaining English words in the Arabic translation have been replaced by several "*" characters. This should allow us to study the impact of non translated words on judges.

Thus, 6,120 sentences were evaluated twice and randomly distributed among 50 different judges. It represents around 245 sentences per judge. Unfortunately, only 11 judges carried out the evaluation, against our expectations. This is also why dry-runs are performed and recruitment modalities have been modified for the evaluation campaign. As a whole, this implies 1,548 evaluated sentences, with around 129 sentences per system.

The number of participants limits the interest of the human evaluation. Even if measuring performance was not our goal, it gives an idea of what we could expect from the evaluation campaign. This is due to the period when the dry-run took place (summer break): judges were contacted but there was a clear lack of motivation (certain judges did start the judgements but stopped when they realized the task was difficult or unpleasant). Results give a BLEU correlation of 98% with adequacy and of 96% with fluency, showing a very good correlation with human metrics, much higher than for other languages (Callison-Burch et al., 2010).

[11] http://translate.google.fr/?hl=fr&tab=wT#
[12] http://www.systran.fr/

The dry-run has been useful to test our protocol, interface, scripts and metrics. Regarding the results, it seems the test corpus is difficult to translate, even for a professional translator. Within the evaluation campaign, the results are expected to be better after deploying the training corpus.

4.2 Evaluation Campaign

The dry-run gave us an idea of the baseline systems' performance and allowed to develop a first evaluation framework for English-to-Arabic. Therefore, we planned an evaluation campaign that aimed at testing systems after their tuning. Training data was provided to improve the systems and all the participants were limited to it.

4.2.1 Participating Systems

Six submissions have been received from four participants: ENSIAS, Sakhr, University of Balamand (UoB), and Columbia University (CU). Only the last one is an external participant, the others being members of the MEDAR consortium. Four submissions from the two MEDAR systems have been made. ENSIAS used a Moses-based system derived from Baseline 2 only trained with the MEDAR corpora. UoB used an improved version of Baseline 1 introducing new functions such as a simple morphological analysis so as to improve prefix processing and handle synonyms in the translation. CU (El Kholy & Habash, 2010) used a language model based on the IRSTLM toolkit (Federico et al., 2008), Moses for training and decoding and the Penn Arabic Treebank tokenization scheme to preprocess the Arabic data, with all the parallel training data but the MEDAR Dry-run that has been used for decoding weight optimization. Sakhr used a morphological analyzer, based on an Arabic lexicon and associated linguistic features, automatic diacritization and rule-based MT. The same two online systems (Google Translate - statistical-based - and Systranet - rule-based) have been used for our needs but their results must be considered carefully since they are not tuned for our data.

Up to 5 submissions were allowed per participant, allowing participants to tune their systems with different parameters as they felt appropriate. The first submission is identified as "primary", the others as "secondary" (not shown in this paper).

One version of the Baseline 1 system has been submitted using all the parallel training data. Three versions of the Baseline 2 system have been submitted, according to the monolingual and parallel training data used, as presented in Table 3.

System	Monolingual	Parallel
Baseline 2-1	All	All
Baseline 2-2	Baseline	LDC
Baseline 2-3	MEDAR+ELRA	MEDAR

Table 3. Training data of the Baseline 2 system.

The MEDAR evaluation campaign has been carried out in July 2010. The schedule was tight for participants but we have been pleased to see all of them have been able to respect the project's deadlines. They had 15 days to train their systems and 5 days to return their translations. In two days, the automatic results were made available. The human evaluation has been postponed to two months later, mainly because of the summer break.

4.3 Analysis of the Training Material

In order to analyse the parallel training data, we split it in two parts (see Table 2): LDC training data consist in Arabic newswires, while MEDAR data contain the climate change domain, the Meedan memory translation containing news data and the UN data, close to diplomatic domain but quite heterogeneous. We then compared the training corpora to the test corpus, particularly focusing on the vocabulary used and the lexicon size. To do so, we computed the number of different English words for both LDC and MEDAR parallel corpora (see Table 4).

Corpus	#Lexicon (En)	Occurrence mean
LDC	27,276	28.5
MEDAR	28,797	91.3
LDC+MEDAR	41,789	81.6
Test	2,444	3.7

Table 4. Statistics on training and test corpora.

Both LDC and MEDAR training corpora are quite similar in terms of number of different words. Half the words are in the two corpora. We also observed that most of the words are not frequent in the corpora and a few words are far more frequent, such as non-content words ('the', 'a', etc.). The means of word occurrence indicate that the LDC parallel corpus is more heterogeneous than the MEDAR one. Indeed, there is more variety of lexicon in the former than in the latter, that is more repetitive. However, the amount of unique words is quite similar: 10,436 for LDC against 10,614 for MEDAR. Likewise, the difference in the number of words that occur between 2 and a hundred times remains stable between the two corpora.

We then observed the out-of-vocabulary of the test corpus according to the training corpora.

Corpus	#Lexicon (En)	Lexicon OoV	#words	Words OoV
Test	2,444	-	8961	-
LDC	27,276	384 (16%)	778K	604 (7%)
MEDAR	28,797	250 (10%)	2,630K	388 (4%)
All	41,789	194 (8%)	3, 409K	306 (3%)

Table 5. Out-of-vocabulary of the test corpus.

A substantial part of the lexicon is unknown to the MT systems when translating the test corpus. Around 16% of the test corpus lexicon is unknown when dealing with the LDC corpus, which is quite important. Figures are still important using the MEDAR training corpus (10%) or the overall training corpus (8%). However, unknown words are less frequent since the percentage of OoV on words is lower than the percentage of OoV on lexicon. The impact is obvious: when 3% of the test corpus is OoV in the training corpus BLEU does not match at least 3% of the n-grams.

4.4 Evaluation Results

10 submissions have been manually evaluated. Each submission contained 390 sentences. Thus, 7,920 sentences were evaluated twice and randomly distributed among 50 different judges, representing around 158 sentences per judge.

To test the agreement among judges, we compute an inter-judge *n*-agreement, for which *n* is the upper difference between two scores of the same segment (Hamon et al., 2008).

Evaluation	n				
	0	1	2	3	4
Fluency	.38	.78	.94	.99	1
Adequacy	.37	.69	.85	.93	1

Table 6. Inter-judge n-agreement [0-1].

Results are similar to previous experiments: for almost 40% of the evaluated sentences, judges give the same scores, which is rather low but shows the difficulty and the subjectivity of the judgements. However, *n*-agreements when *n>0* are high and prove the evaluation is reliable.

4.5 Results

The results of the evaluation are shown in Table 7.

System	BLEU [%]	Ade. [1-5]	Flu. [1-5]
Human ref. 1	*69.7*	*4.34±.07*	*4.11±.08*
Google Translate	*20.8*	*3.45±.10*	*3.49±.08*
Sakhr	15.2	3.27±.09	3.26±.08
CU	12.6	3.07±.10	3.30±.09
Baseline 2-3	6.5	2.03±.09	1.74±.08
Baseline 2-2	6.3	2.16±.10	1.83±.08
Baseline 2-1	6.1	-	-
Baseline 1-1	6.1	2.34±.09	2.12±.09
ENSIAS	5.6	1.77±.07	1.41±.05
UoB	3.8	2.17±.09	1.92±.08
Systranet	*2.0*	*2.23±.08*	*2.05±.08*

Table 7. Results of the MEDAR evaluation campaign.

Automatic results are ranked differently to human ones in the second part of the table. The order of the baseline systems is reversed. However, translations are very close and these differences are not significant enough to draw any conclusion. Regarding low UoB results, judges may be surely influenced by the number of untranslated English words in the Arabic translation.

Human results show a clear hierarchy among the translations. First, human translation obtains high results, but not as high as expected. As for other campaigns in the MT domain, translations are not perfect, judgements may differ and, to a certain extent, comparing two human translations means testing the agreement between their translators. Google Translate, Sakhr and CU results are all above 3 points in both fluency and adequacy. Their outputs are rather understandable translations. Results from Systranet, UoB and the three MEDAR MT systems are under average providing translations difficult to understand. MEDAR MT systems get higher results when they use the overall parallel training corpus, which is not really surprising. However, results are higher when using the LDC parallel training corpus instead of the MEDAR one, even if OoV is bigger when using the former. Here, the monolingual corpus seems to have deteriorated the quality of the translations.

Furthermore, looking at the judgements in detail, we identified the following five general problems the MT systems may have to address.

Missing lexicon entries: OoV words are either kept in English or transliterated. The former affects the quality perceived by judges. The latter is either hardly understandable by judges – because of a specific vocabulary not close to their knowledge – or contains one or several Latin characters that causes definitely the incomprehension of what is said. It also seems that some good transliterations are not well scored by judges due to either lack of knowledge or another existing word in Arabic.

Src.	High levels of arsenic in seawater can enable the toxin to enter the food chain.
Ref.	المستويات العالية من الزرنيخ في مياه البحر يمكنها أن تسمح للسم بالتسلل إلى سلسلة الغذاء.
MT	يمكن أن seawater في arsenicارتفاع معدلات الإصابة دخول الغذاء تقييد.toxinمن

Table 8. Example of unknown words (flu.=1; ade.=1).

Compound words: they can be either considered as a named entity or be translated as independent terms. Therefore, the meaning of the translation is strongly modified.

Src.	Adapt **land use regulations** to the potential rise in sea level, by **increasing the minimum clear distance** required between buildings and shoreline.
Ref.	تكييف أنظمة استخدام الأراضي إلى احتمالات ارتفاع في مستوى مياه البحر ، وذلك بزيادة الحد الأدنى المطلوب مسافة واضحة بين المباني . و
MT	أنْ تُكيف قوانين استخدام الأراضي مع الارتفاع المحتمل في مستوى سطح البحر، بواسطة زيادة الحد الأدنى للمسافة الفاصلة بين المباني والشاطئ.

Table 9. Example of compound word (flu.=5; ade.=1): *clear* is translated as a word instead of *clear distance*.

Complex sentence translation (comprising coordinated structures, subordinated structures or sentences): syntactic issues arise when translating complex sentences. Complex sentences may not be identified as such or segments may not be split correctly. This implies that translation does not focus on the correct meaning. This is particularly

so with long sentences. Generally speaking, the longer the sentence, the bigger the chance to have syntactic issues due to the weak identification of the sentence construction. This is the case for Baseline 1 and 2, but even systems such as Sakhr or Google Translate are concerned.

Src.	The calculates future global aviation emissions of carbon dioxide and NOx from air traffic under four of the IPCC/SRES (Intergovernmental Panel on Climate Change/Special Report on Emissions Scenarios) scenarios:
Ref.	إنهم يحسبون انبعاثات الطيران العالمي مستقبَلاً من ثاني أكسيد الكربون وأكاسيد النِتْريكْ بفِعْل الحركة الجوية حسب أربعة سيناريوهات (المجلس الحكومي الدولي للتغير المناخي/ التقرير الخاص عن سيناريوهات الانبعاثات):
MT	بمستقبل الطيران العالمية مؤاتيا لانبعاثات ثاني أكسيد الكبريت ومن حركة الملاحة الجوية في إطار أربع منها (. .) على الفريق الحكومي الدولي المعنى بتغير المناخ تقريرا خاصا عن (انبعاثات) : سيناريوات سيناريوات

Table 10. Example of complex sentence (flu.=3; ade.=1): parts of the sentence are badly translated; proper names are missing; several dependent clauses are hard to split.

Wrong syntactic analysis or lemmatisation: some words are not well tagged (e.g. as a noun instead of a gerund), causing a mistranslation. Both fluency and adequacy are therefore hard to follow.

Src.	They discovered that seawater alters the chemistry of goethite, where low pH levels in the water create a positive change on the surface of goethite sediments, making them attractive to the negatively charged arsenic.
Ref.	لقد اكتشفوا أن ماء البحر يبدل في كِمْياء الجُويثايت، حيث المنخفضة في المياه شحنة موجبة على pH توجِد مستويات سطح رواسب الجُويثايت، مما يجعلها جاذبة للزرنيخ سالب الشحنة.
MT	مِستري ل eهcرس ب ير َلتeوَتeو اكتشفت أن س , حيث منخفضة ةح مستويات المياه في تهيئة eتهتeعَ نتس , مما eمِمeسeتهتeإيجابية تغيير في السطحية ل عَ .eنeيجعل ها جاذبية ب شكل سلبي متهمةَ رس

Table 11. Example of wrong lemmatization (flu.=1; ade.=1): words are cut, not translated but transliterated.

Named entity translation: many are either not translated or not well transliterated. This is mainly due to missing vocabulary in the training data. This causes a big drop in fluency (when the translation is poor, missing named entities do not help rebuild the sentence correctly) and less often in adequacy. Indeed, missing named entities do not imply the meaning is hard to find: we can understand that somebody did something without knowing who.

Src.	Hemlock Semiconductor just started building a polysilicon plant in Tennessee.
Ref.	وشرعت هِيمْلُوك لأشباه الموصلات للتوّ في بناء مصنع للبُولي سِيلِيكون في تِينِيسِي.
MT	hemlock semiconductor مسلّح لبناء polysilicon اليرموك في.

Table 12. Example of named entities not translated with wrong word order (flu.=1; ade.=1).

For the baseline systems in particular, we observed typical errors according to the level of fluency score. When many words are not translated, especially named entities, the fluency score is often put at its lowest level. A fluency score of 2 (second lowest level) is generally linked to a wrong generation in the target language, showing the limits of the language model. Moreover, the Arabic morphology is not well respected: many suffixes or prefixes are not agglutinated properly. Fluency scores of 3 and 4 correspond to different levels of problems regarding the way the semantics is rendered into the syntax. This is often the case in complicated English source sentences with over 3 or 4 connected clauses as, for instance, number and gender are badly rendered in the Arabic syntax.

In the same way, adequacy scores are affected by typical errors, such as those in fluency. Due to the pretty low translation level, a non fluent translation also affects the understanding of the meaning. Moreover, numbers in numerical characters may be an issue for MT systems, wrongly translating the corresponding term. For instance, “2 actions” translated into “2 years”: the translation model is confused by a mistranslation in the training data.

There is also a number of sentences that are correct in terms of fluency (i.e. the language model and the reordering are working) but that obtain a low adequacy (i.e. the decoding or the translation model are low).

The judgements allow us to evaluate the efficiency of BLEU, that worked well, but not perfectly: Pearson correlation coefficients on BLEU are of 0.78 for fluency and of 0.90 for adequacy.

5 Lessons Learnt and Further Work

Using training data seemed to improve the scores, at least by one point of BLEU, but the performance within MEDAR is still too low compared to current systems using similar approaches for other languages. A number of open issues have to be tackled in order to improve such performance.

The size of training data must be increased, with a higher quality that fits the vocabulary of the test corpus. This can be achieved by importing data from several domains and using a large range of lexica. Working with OoV can be a complex task, but solutions exist, such as in (Habash, 2008).

Another solution is to incorporate more tools managing the specific features of Arabic, especially regarding preprocessing. The generation post-processing is also essential and requires more work for sentence reconstruction, like looking at gender or number. Likewise, Moses should be improved for Arabic, using word reordering for alignment, syntactic analysis for preprocessing, segmentation and morphological decomposition, word alignment, etc.

Regarding evaluation, we need to ensure scoring metrics are appropriate to assess Arabic output.

The goal of MEDAR was not to provide an advanced, free, open source, system for MT from English to Arabic but rather to initiate activities in that direction and rise interest. We felt the best approach was to offer an evaluation framework and considered that, despite all MT R&D efforts, most of the work is on Arabic as a source language. In that context, MEDAR allowed the community to benefit from the evaluation data developed and the organization of an evaluation campaign.

Despite the low performance achieved by several systems based or derived from Moses, MEDAR is happy to offer these packages to the HLT community. They contain the 2 baseline MT systems, the test and masking corpora of the dry-run and the evaluation campaign together with their respective 4 reference translations, and the MEDAR monolingual and parallel training data. The LDC data may be obtained through LDC.

Furthermore, by offering such a package to the researchers and students, we hope to boost activities on MT for English-to-Arabic and further considering Arabic as the target language.

Acknowledgments

This work was supported by the MEDAR project (grant number ICT-214602). We would like to thank Bente Maegaard, Chafic Mokbel, Ossama Emam, Sara Noeman, Dorte Haltrup Hansen, Mohamed Attia, Mossab Al Hunaity and Abdelhak Mouradi for their active collaboration. We are also very grateful to all the evaluation participants and to Victoria Arranz for her feedback on the article.

References

Callison-Burch C., Koehn P., Monz C., Peterson K., Przybocki M., and Zaidan O.F. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR (WMT'10). ACL, Morristown, NJ, USA, pp. 17-53.

El Kholy A. and Habash N. 2010 Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In Proceedings of *Language Resources and Evaluation Conference (LREC) Workshop on Semitic Languages*, Malta.

Ghaoui A., Yvon F., Mokbel C. and Chollet G. 2005. On the Use of Morphological Constraints in N-gram Statistical Language Model. In Proceedings of *Interspeech 2005*, pp. 1281-1284.

Federico M., Bertoldi N. and Cettolo M. 2008. IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. In Proceedings of *Interspeech 2008*, pp. 1618-1621.

Habash, N. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In proceedings of *the 46th Annual Meeting of ACL*. Columbus, Ohio.

Hamon O., Mostefa D. and Arranz V. 2008. Diagnosing Human Judgments in MT Evaluation : an Example based on the Spanish Language. In Proceedings of *MATMT*, San Sebastian, Spain, pp. 19-26.

Hamon O. and Choukri K. 2010. Evaluation methodology and results, Mediterranean Arabic Language and Speech Technology (MEDAR) project. Deliverable D5.3.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertordi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007*, pp. 177-180.

Och, F.J. and Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, vol. 29, pp. 19-51.

Papineni K., Roukos S., Ward T., and Wei-Jing Z. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM Research Division, Thomas J. Watson Research Center.

Stockle, A. 2002. SRILM - An Extensible Language Modeling Toolkit, In Proceedings of *International Conference on Spoken Language Processing*.