MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable 5.3

Evaluation of MT Systems for Arabic

Author: Khalid Choukri, ELDA, Olivier Hamon, ELDA

September 2010

MEDAR partners

- **University of Copenhagen:** Centre for Language Technology, Denmark (coordinator)
- **ELDA,** Evaluations and Language resources Distribution Agency , France
- **University of Balamand:** Research Council - Speech and Image Research Group (SIR), Lebanon
- **Amman University:** Faculty of Information Technology, Jordan
- **University of Utrecht:** Utrecht Institute of Linguistics OTS, the Netherlands
- **Research and Innovation Centre "Athena":** ILSP, Institute for Language and Speech Processing, Greece
- **RDI,** The Engineering Company for the Development of Computer Systems, Egypt
- **Birzeit University:** Center for Continuing Education, West Bank and Gaza Strip
- **University Mohammed V Soussi:** Ecole Nationale Supérieure d'Informatique Analyse des Systèmes, Morocco
- **CEA,** Commissariat à l'Energie Atomique: CEA-LIST/LIC2M, Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue, France
- **CNRS,** Centre National de la Recherche Scientifique, Laboratoire LLACAN - UMR 8135 du CNRS, Langage, langues et cultures d'Afrique Noire, France
- **The Open University:** Computing Department, Maths & Computing Faculty, The United Kingdom
- **Université Lumière Lyon2:** Groupe SILAT, France
- **IBM** International Business Machines WTC - Egypt Branch, Egypt
- **Sakhr** Software Company, Egypt

The MEDAR project is supported by the ICT programme

© The authors and MEDAR, c/o Center for Sprogteknologi, University of Copenhagen, September 2010, http://www.medar.info, email: nemlar@hum.ku.dk

CONTENT

1. Executive Summary

This report deals with the evaluation methodology and results of the two MEDAR evaluation campaigns. The context is the evaluation of MT systems for English-to-Arabic direction. The very first goal is to identify the performance level of the MEDAR baseline systems developed within the WP5.

The evaluation is conducted in two phases. Phase 1 aiming at setting some basic facts about state of the art for MT on English to Arabic while the second one aimed at collecting enough data to better train and tune the systems and assess the improvements made.

The report describes the data used and their formats, the preparation of the campaigns as well as the results of the systems. MEDAR allowed the community to benefit from the evaluation data developed and the evaluation organization in participating to the two evaluation campaigns. Thus, several external systems have been evaluated in addition to the MEDAR baseline systems.

A couple of online translation systems have been used to compare with the results submitted by our participants. Interpretations of such results have to be made with a lot of care as these systems have not been tuned to our data.

Finally, the report gives several recommendations in MT evaluation for English-to-Arabic direction in terms of technologies and in terms of resources.

2. Objectives of MEDAR Evaluation Campaigns

When dealing with Arabic, most of the evaluation campaigns or MT systems consider the Arabic-to-English direction only. One of the major goals of MEDAR is to experiment and develop the research around the English-to-Arabic direction.

Therefore, MEDAR evaluation campaigns target several objectives:

- Developing a framework for the evaluation of English-to-Arabic MT systems;
- Producing data for MT training;
- Producing data for MT evaluation;
- Developing a baseline with background from existing open source tools;
- Evaluating MEDAR MT baseline systems;
- Ranking MEDAR baseline MT systems regarding other MT systems;
- Creating and federating a new community around the MT English-to-Arabic theme;
- Making available a package containing the full set of resources and tools from MEDAR.

3. First MEDAR Evaluation Campaign

3.1. Evaluation Corpus

3.1.1. Material and Preparation

To proceed with the test of the systems, a test corpus must be built, as well as a masking corpus. The test corpus allows scoring the systems against reference translations, which are made by human high quality translations of the test corpus. The "masking" corpus is much larger and is used to hide the test corpus to the participants and thus, participants should not be able to identify the test corpus. After receiving the submissions from participants, only the part corresponding to the test corpus is kept.

Input data are English texts coming from a specific domain (climate change). They are composed of about 210,000 running words, from which 10,000 words are used as a test corpus, the rest being the "masking" corpus.

The overall evaluation data has been built as follows:

1. The 210,000 words in the evaluation data have been collected from many different websites whose material discusses the topic of Climate Change.
2. Part of this test data, a test corpus of about 10,000 words, has been selected to evaluate the MT systems.
3. The remaining words are used as a masking corpus in order to keep unknown the part that will serve as the test corpus and ensure that no post-processing is done by participants (post-editing, corrections, etc.).
4. The 10,000 words of the test corpus have been translated four times by four different translation teams (one translation per translator). Specific guidelines were produced, and provided to the translation agencies in order to control the quality of their produced translations. Likewise, specific validation guidelines were also produced for validating these translations.

3.1.2. Translation Guidelines

3.1.2.1. Goal

The goal of the translation guidelines is to support the production of a corpus for the evaluation of machine translation systems. The objective of the work is thus to produce high-quality bilingual data, by translation professionals and to ensure that such outcome represent the target against which to compare the MT systems.

3.1.2.2. The Translation Team

Each translation team is used to translate all of the source language data. Such a team is composed of:

1. Several bilingual translators, native speakers of the target language of the data (Arabic).

2. A bilingual target native speaker who proofreads and edits the output of the translators. He/She is also in charge of the homogenization of the whole test corpus, especially regarding the vocabulary and terminology within the text.

Notice that the translations are systematically finalized and checked by a target native speaker. The translation team does not change during the course of translation, and the team composition is fully documented. The documentation includes:

- The name (or pseudonym), native language, second languages, age and years of translation experience of the translator(s).
- The order of processing (i.e. the name of the person who performs the first pass, second pass, etc.), together with the names of the files handled.
- The name and version number of any translation system or translation memory used.
- A description of any additional quality control procedures or other relevant parameters or factors that affect the translation.

3.1.2.3. Material

Data are monolingual texts coming from a specific domain and have an average length of twenty words per sentence. They may come from website and other internet sources. Thus, the translators are requested not to use any related translated data that may exist on the Internet. The translation team should not use these sources (neither English nor Arabic parallel pages) for their translation. The use of these websites should be avoided.

The translated file is rendered in XML format, UTF-8 encoded, so as to preserve the original structure.

3.1.2.4. Translation Quality

Translation agencies used their best practices to produce the MEDAR translations. While we trust that each translation agency has its own mechanisms of quality control, we have specific guidelines so that all translations share a common ground. These are:

1. The target translation must be faithful to the original source text in terms of meaning and style. When the source text is a press release, the translation should be written in a journalistic style, thus respecting the document style. The translation should mirror the original meaning as much as possible without sacrificing grammaticality, fluency and naturalness.
2. The tone and register of the language should be respected. For instance, if the text shows an angry or uneasy speaker in the source language, this state of mind should be also expressed in the target language, conveying the same tone.
3. The same applies for the general ”politeness” and ”formality” register of the source text. Both translators and proofreaders should bear in mind the ”politeness” standards of the target language.
4. The translation should be as factual as possible, trying to keep the exact information conveyed by the source text, without changing the meaning and without adding/removing information. For example, if the original text uses ”Obama” to refer to the U.S.A. President, the translation should not be rendered as ”President Obama”, ”Mister Obama”, etc.

5. No bracketed words, phrases or other annotation should be added to the translation as an explanation or aid to understanding.

6. The translation should entail the same cultural assumptions as the original text, and no implicit reference should be made explicit by the translator.

7. The order of consecutive segments must not be altered, not even for stylistic reasons, i.e. the contents of segments N and N+1 must not be swapped in the translation.

8. Capitalization and punctuation are language dependent. This means that translators should follow the standards from the target language and apply their rules even if these may not coincide with those of the source document.

9. Regarding neologisms and unknown words: if it is possible to understand the intention/gist of the source text, then the translation should be either the correct form of the word (for unknown words) or a new word corresponding to the source derivation (for neologisms). If the translator has no preexisting knowledge on how to translate a word, (s)he is expected to consult standard sources, such as dictionaries, translation forums, etc.

10. Regarding proper names, whenever possible, these should be translated following conventional practices in the target language. For instance, in the case of Arabic, this may imply providing a different translation from that suggested in Modern Arabic. The order of the family name and first name presentation should be preserved as that of the source file. As with neologisms, when lacking knowledge on the word to translate, translators are expected to consult standard resources.

11. The format of entities like dates and numbers in general must remain the same in the translated document.

12. Idioms and colloquial expressions are particularly hard to translate. If a similar expression exists in the target language, it should be used. However, if there is no direct translation into the target language, translators should try to preserve the meaning of the source-language expression but convey it in as natural and fluent a target-language expression as possible.

13. The normalization and revision of the whole corpus will be done in terms of terminology used, as well as orthographic consistency, style and register. For consistency purposes, the proofreading of the full corpus will be done by a target native speaker.

3.1.3. Validation Guidelines

The goal of the validation guidelines is to provide a methodology for validating the translations produced. These translations are validated by a team of expert validators. Validation is done according to the translation guidelines.

3.1.3.1. Procedure

Once finalized by the translation agencies, translations are validated. Validation follows the specific criteria described below.

Resulting translations are divided into *accepted* and *rejected*. An accepted translation is kept, while a rejected translation is sent back to the translation agency with a validation report and the errors found. A delay is agreed upon for the return of a new translation. As the validation procedure is carried out on a sample of each translation, the new translation to be provided by the translation agency must not be a corrected version of this sample only, but of the full file.

The validation of the data consists of both an automatic and a manual procedure.

3.1.3.2. Automatic validation

An automatic validation is provided when a translation is received from the translation agency. If numerous and irrefutable errors are found, the translation is immediately sent back to the translation agency.

The following issues are considered in this automatic validation:

- A spell checker checks the translation automatically. If necessary, the spell checker is adapted to the corpus lexicon. The errors found are considered as lexical errors described in the scheme below, and are included in the final validation report.
- The format of the corpus is automatically validated too, checking whether the specifications established in the translation guidelines have been followed. The translation might be sent back to the translation agency if the number of errors found is above a threshold.

3.1.3.3. Validation by human experts

Regarding manual validation, this takes place over a selected sample of data. The guidelines detailed here are used for the selection of the material to be validated as well as for its validation.

For each delivery, a random subset of sentences of the test corpus is selected at ELDA, until the number of words adds up to about 5% of the source text (considering full sentences) translated by a single translator. Then, the validation corpus is supplied to the validators (one per translation) containing both source and target texts.

The validation task consists in proofreading the texts and whenever a problematic point arises:

- Label the problematic sentence (with a label from the list of problems detailed in the table further down in Point 4);
- Propose a correction/improvement, if possible and/or a short explanation of the error found.

The task of the validator is to evaluate if the translation is of good quality, not redo it, as when aiming to produce a final version of a document for publication. Such revision/correction is the task of the translation agency. However, since we are evaluating the quality of the data we certainly need validators to provide arguments (some corrections, comments) to prove the validator's criteria/decisions.

The following technical issues should be taken into account:

- Files to be validated are provided to validators in text format (or Microsoft Office Word, if required). Validators are expected to submit their files respecting this original format.

- The sentences to be validated look as follows :

  source sentence

  translated sentence

  blank line

- Corrections and notifications of errors are provided per sentence. If no remark or correction is to be provided by the validator, this format remains the same. However, if a segment contains an error, then a new line is inserted starting with "#" right after the segment. After the "#" follows the type of error (5 categories, according to the scheme described below), together with the correction or indication of the error itself. The resulting format is as follows:

  source sentence

  translated sentence

  # error type + correction or indication of the error

  blank line

- In the case of multiple errors, each error is on a new line starting with "#". Notifications and remarks should be made in English.

- To ensure consistency from one validator to another, the following system has been adopted for grading translations. Validators use the following types/labels (whenever possible) to tag translation errors: Syntactic, Lexical, Poor usage of target language, Punctuation.

Syntactic errors	are those found in grammatical categories. These comprise errors such as problems with verb tense, coreference and inflection. Furthermore, syntactic errors are also those where there has been a misinterpretation of the grammatical relationships among the words of the original text. Examples of syntactic errors are, for instance, translating an object as a subject, making an adjective modify a verb, attaching a relative pronoun or prepositional phrase to the wrong noun.
Lexical errors	comprise omitted words or wrong choice of lexical item (word), due to misinterpretation or mistranslation.
Poor usage of target language	means awkward, unidiomatic usage of the target language and failure to use commonly recognized titles and terms.
Punctuation errors:	Punctuation should also follow the standards/conventions of the target language, even if the source language is not correctly punctuated.

Table 1: Type of Errors

It is essential that the given translation receives the "benefit of the doubt". Only clear errors should be indicated.

When several translations are produced for a same source text, these are validated separately, each of them going through the same validation procedure described above. However, serious errors (syntactic and lexical) detected in either one of the translated texts are also verified in the other translations in order to avoid the proliferation of problematic cases. This verification among the different translations is based on the results/findings of the validations.

3.1.3.4. Validation criteria

A validation score is computed as the sum of errors found by validators, according to both the number and type of errors found. If the score is above an allowed threshold, the translation is rejected and, thus sent back to the translation agency for correction. A complete revision is required and not only for the sub-set randomly selected for validation.

3.1.3.5. Validation report

When a new translation is validated, a validation report is produced, allowing the follow-up of the translation procedure and the interaction with the translation agency.

3.1.4. Formats

The input corpus is encoded in XML and UTF-8 and contains documents identified with a *docid* attribute and a genre code. Each sentence within a document is tagged and identified

with an *id* attribute. The format specifications of the DTD and an example are given in Annex A.

3.2. Outputs

The output files are to be returned back in XML format, UTF-8 encoded, so as to preserve the original structure. A *sysid* attribute is added to the DOC tag. The DTD and an example are given in Annexe B.

3.3. Participating Systems

In MEDAR, two baseline SMT systems have been used. They are developed by the University of Balamand ("Baseline 1") and IBM in partnership with DCU ("Baseline 2") on the basis of Moses[1]. Moses is an open-source statistical machine translation system and the two baseline systems have been adapted so as to translate from English to Arabic.

Baseline 1 provided a Moses system improved with language models to be installed by all MEDAR partners.

Baseline 2 has been built according to the procedure below:

1. Downloading and installing of the required packages for Alignment (GIZA++v2), Language modeling (SRILM), Arabic and English tokenization (AMIRA-1.0 for Arabic, and opennlp-tools-1.4.3 for English), Moses package for block extraction and beam search decoding.

2. Packaging a DVD of all binaries required and full packages that need to be installed locally as well as a sample of LDC data to be used in demo training-decoding experiment. Full documentation is included in that package so as to install tools as well as setting all the environment variables required by the package.

3. Preparing a sample end-to-end script which performs:

    a. Tokenization for the required English-Arabic data;

    b. Building of a Language model using the target language training data (Arabic);

    c. Application of a statistical alignment on the parallel corpus using Giza++;

    d. Building of a phrase model;

    e. Decoding the sample test data using "moses" decoder.

Furthermore, the evaluation campaign was open to external participants and participants from the MEDAR consortium. Therefore, a promotion of the campaign has been made through several procedures: mailing lists, networking, personal contacts, conferences, etc. Four participants (one external and three from MEDAR) replied and five submissions have been made. The lack of participation may be explained by the short delay between the start of the

[1] http://www.statmt.org/moses/

campaign and the scoring. However, it also may be due to the lack of existing English-to-Arabic systems in the field. For this first MEDAR evaluation, five submissions have been received, anonymized and renamed as “System A” to “System E”

Finally, for comparison purposes, two online systems have been used in this evaluation: Google Translate[2] and Systranet[3]. Their results must be considered carefully since they are not really participating systems.

3.4. Training and Development

There was no training or development phase planned for the first MEDAR evaluation campaign, therefore no data is provided to participants. The two MEDAR baseline systems have not been specifically trained, a very basic data set has been used (this is a small corpus included in each package).

Participants were free to use any kind of data they could obtain. Therefore, systems are not directly comparable. Their results are presented hereafter just to give an idea of their relative performance. They remain anonymised.

2 http://translate.google.fr/?hl=fr&tab=wT#

3 http://www.systran.fr/

Evaluation Schedule

The schedule of the first MEDAR evaluation campaign was specified as follows:

January 19, 2010	Evaluation data are sent to participants
January 29, 2010	Deadline for sending back translations
February 03, 2010	Preliminary automatic results
February 07, 2010	Final automatic results after checking

Table 2: Schedule of the first MEDAR evaluation campaign.

3.5. Results

3.5.1. Scoring Tools

Automatic scoring is done using BLEU, BLEU/NIST and mWER metrics at ELDA.

BLEU, which stands for BiLingual Evaluation Understudy, counts the number of word sequences (n-grams) in a sentence to be evaluated, which are common with one or more reference translations. A translation is considered better if it shares a larger number of n-grams with the reference translations. In addition, BLEU applies a penalty to those translations whose length significantly differs from that of the reference translations.

BLEU/NIST, is a variant metric of BLEU, from NIST (*National Institute of Standards and Technology)*, which applies different weight for the n-grams, functions of information gain and length penalty.

mWER, Multi reference Word Error Rate, computes the percentage of words which are to be inserted, deleted or substituted in the translated sentence in order to obtain the reference sentence.

The higher BLEU and BLEU/NIST are, the better our system is (measure of performance); the lower mWER is, the better our system is (measure of error rate).

3.5.2. Automatic (Anonymised) Results

Results have been automatically computed against four references. To compare to what a human translator can produce and put into perspective the results of the automatic systems, the results of one reference translation (*Human reference 1*) is presented below, comparing it against the three other reference translations. Results are shown in Table 3.

System	BLEU [%]	BLEU/NIST [values]	mWER [%]
Human reference 1	*56.34*	*11.00*	*27.50*
Google Translate	*20.31*	*7.02*	*67.67*
System A	16.56	6.32	66.47
System B	11.66	4.78	73.25
System C	11.21	4.97	76.35
System D	5.70	3.88	78.95
System E	5.87	3.54	77.83
Baseline 1	5.08	3.73	80.96
Baseline 2	4.47	3.64	85.81
Systranet	*2.11*	*2.27*	*106.60*

Table 3: Anonymised results of the first MEDAR evaluation campaign.

3.6. Discussion

For this initial campaign no training data was provided to the participants. They were free to use any kind of data they could. The automatic measures showed quite a modest performance at that point. This campaign may be considered as a dry run, so as to test the protocol and the organization and establish the baseline instead of testing the systems objectively.

Therefore, the low scores should be put into perspective. The vocabulary of the test corpus is from a specific domain that is harder to process by the systems. Moreover, the results of one of the human reference translation ("Human reference 1") compared to the three other reference translations are lower than we could expect. Therefore, the test corpus seems difficult for translation, even for a professional translator. On this basis, the results are not as bad as they look.

Finally, we can argue that BLEU or any current automatic metrics may be not adapted to process Arabic data. This could particularly relate to the agglutination of words.

Within the second evaluation campaign, the results are expected to be better after deploying the large training corpus.

4. Second MEDAR Evaluation Campaign

The first evaluation campaign gave an idea of the baseline systems' performance and permitted to develop a first evaluation framework for English-to-Arabic. To go further, we planned a second evaluation campaign that aims to test systems after tuning. Therefore, training data was provided to improve the systems.

4.1. Training

4.1.1. Training Conditions

Two training conditions are implemented in this second MEDAR evaluation campaign: Constrained Training and Unconstrained Training. Participants were asked to participate at least under the first condition.

4.1.1.1. Constrained Condition

In the Constrained Condition, only the data provided by MEDAR can be used for the MT system training. This only refers to Language Resources, and not to tools used by systems. This training condition covers both parallel and monolingual data.

4.1.1.2. Unconstrained Condition

In the Unconstrained Condition, there is no restriction with respect to the data that may be used to train the MT systems. This training condition covers both parallel and monolingual data.

4.1.2. Material (Constrained Condition)

The training data allowed by MEDAR in the Constrained Condition are either parallel data or monolingual data. Parts of the data are provided by LDC which has kindly shared some of the data from its catalogue for the purpose of the evaluation only. Most of the data are available either for R&D (i.e. data produced within MEDAR) or for the MEDAR evaluation purposes (i.e. data from catalogues) only due to copyright constraints. Other data sets are from the ELRA catalogue. Finally, other sets have been collected by the MEDAR consortium.

4.1.2.1. Arabic Monolingual Data

Resources from MEDAR are labelled as Mnnnn; Resources from ELRA or LDC are identified by their respective Unique Identifiers.

Name	Id	Size [words]	Availability
Islamonline	M0001	20M	R&D only
Wikipedia	M0002	31M	R&D only
Wikibooks	M0003	1M	R&D only
Wikinews	M0004	129M	R&D only
Wikiquote	M0005	144M	R&D only
Wikisource	M0006	69M	R&D only
An-Nahar	ELRA-W0027	113M	MEDAR Eval. only
Al-Hayat	ELRA-W0030	38M	MEDAR Eval. only
LMD	ELRA-W0036	475K	MEDAR Eval. only
NEMLAR	ELRA-W0042	494K	MEDAR Eval. only
Arabic Gigaword 4th Ed.	LDC2009T30	2GB	MEDAR Eval. only

Table 4: Monolingual data used for training (Eval.: Evaluation).

4.1.2.2. Parallel Data

The parallel resources packaged within MEDAR are labelled as Medar_Eval1 and MPnnnn; Resources from ELRA or LDC are identified by their respective Unique Identifiers.

Name	Id	Size [words]	Availability
MEDAR Eval. Camp. 1	Medar_Eval1	10K	R&D only
Meedan	MP0001	426K	R&D only
UN	MP0002	2,7M	R&D only
Multiple-Trans. Ar. Part 1	LDC2003T18	23K	MEDAR Eval. only
Ar. News Trans. Text Part 1	LDC2004T17	441K	MEDAR Eval. only
Multiple-Trans. Ar. Part 2	LDC2005T05	15K	MEDAR Eval. only

Table 5: Parallel data used for training (Eval.: Evaluation).

4.1.3. Preparation

4.1.3.1. Monolingual data

Three sources have been used to produce the MEDAR monolingual corpus. ELRA (*Id=W00XX*) and LDC (*Id=LDCXXXXXXX*) corpora are coming from their respective catalogues. Data have been transformed so as to be compliant with the format (in particular its

DTD). No other action has been done (cleaning, selection, etc.) since the content complied with what we were looking for: cleaned data without garbage.

MEDAR corpora have been produced within the project. It consists of 6 corpora coming either from the IslamOnline website or "Wiki" websites (Wikipedia, WikiBooks, WikiQuote, WikiSource). Data from IslamOnline has been crawled, cleaned and formatted according to the MEDAR requirements. Wiki raw data has been downloaded from Wikipedia, then formatted according to the MEDAR DTD; no further cleaning has been made, the data being provided without garbage content by the "Database Dump" of Wikipedia[4].

For all these resources, IPR issues have been cleared to allow their use within these evaluations but also as parts of the MEDAR evaluation package, an important exit strategy of the project.

4.1.3.2. Parallel data

Three sources have been used to produce the MEDAR parallel corpus. 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 constituted using the corpus developed during the first evaluation campaign. It consisted of the test corpus and the four "references" translations, formatted into four parallel corpora of 10K words.

Two parallel corpora have been selected from already existing data: Meedan translation memory and UN corpus originally available from http://www.uncorpora.org.

Again, for all these resources, IPR issues have been cleared to allow their use within these evaluations but also as parts of the MEDAR evaluation package.

4.2. Evaluation corpus

4.2.1. Material

Evaluation data are English texts coming from the same domain as for the first campaign (climate change). They are composed of about 40,000 words, from which 10,000 words are used as a test corpus and the other 30,000 words as a masking corpus. Input and output formats are the same as for the first MEDAR evaluation campaign. Moreover, the preparation of the evaluation data has been done in the same way as for the first MEDAR evaluation campaign.

4.3. Submissions

Several submissions were allowed per participant, up to a maximum of 5. If more than one output per system is submitted, one of them must be identified as the "primary" submission. Others are considered as "secondary" submissions.

---

4 http://download.wikipedia.org/backup-index.html

The idea behind multiple submissions is to allow participants to tune their systems with different parameters if they feel this is appropriate in this context of R&D evaluations.

A *sysid* attribute identifies the organization, the condition and the system of the submission. For instance, if the organization ORG submits one primary submission and two secondary submissions, then 3 files will be sent with the following *sysid*: ORG-PRIMARY, ORG-SECONDARY1, ORG-SECONDARY2.

4.4. Participating Systems

4.4.1. Overview

As for the first MEDAR evaluation campaign, two online systems have been used in this evaluation: Google Translate and Systranet. Six submissions have also been made by four participants: ENSIAS, Sakhr, the University of Balamand, and the University of Columbia. Only the latter is an external participant, the other participants being members of the MEDAR consortium.

Six submissions for the two MEDAR baseline systems have been made. They are developed by the University of Balamand ("Baseline 1") and IBM in partnership with DCU ("Baseline 2"), on the basis of Moses.

All the submissions have been done within the Constrained Condition. Hereafter are the descriptions of the systems as provided by the participants.

Among the participating systems and to the best of our knowledge, one is a rule-based MT system while the others are statistical-based MT sytems. Among the online systems, Google Translate is a statistical MT system and Systranet is a Rule-Based system. This should be taken into consideration to the interpretation of the results: it is well-known that the BLEU metric, and to a certain extent the other automatic metrics, penalize rule-based MT systems against statistical MT systems.

4.4.2. Description

The descriptions below are provided by the participants to the second MEDAR evaluation campaign.

4.4.2.1. ENSIAS

ENSIAS used a Moses-based system derived from the MEDAR Baseline 2, without the tuning part. To build the translation tables and the language model, the system has been trained with the Medar_Eval1, MP0001 and MP0002 corpora.

4.4.2.2. Sakhr

Sakhr is an active player on the commercial market and have been offering MT systems and services for more than a decade.

The first component of the Sakhr (E>A) MT system is the English Stemmer. The stemmer is based on an English lexicon that contains valid stems along with their part of speech (POS), syntactic features (e.g. transitivity, agreement, pre-terminals), and semantic features (e.g. senses, taxonomies). For each English token, the analyzer generates a list of valid analyses. The correct analysis is determined according to context, using additional information from

databases of proper names, idioms, adverbs, and word collocations, as well as grammar rules that use all information contained in the lexicon.

The second step in the Sakhr system is machine translation itself. This process uses the information from the components described above to disambiguate the English words, and assign feature values to them. This input is used, together with English grammar rules to produce a full parse of the source sentence. Transfer rules, and an English-to-Arabic lexicon are then used to transform the English parse tree to Arabic. A generation step is then applied to the output sentence in order to make it more grammatical. This step applies agreement rules among other things. The last step is to make the output more fluent by applying surface transform rules, and a database of Arabic expressions.

4.4.2.3. University of Balamand

The University of Balamand used an improved version of the MEDAR baseline 1 system. New functions have been introduced regarding the baseline system:

- Simple morphological analysis so as to improve the prefix processing;
- Consideration of synonyms in the translation.

4.4.2.4. University of Columbia

All of the training data are from the provided constrained list in the evaluation plan. The system uses an English-Arabic parallel corpus of about 114K sentences and 4 million words for translation model training data. The parallel text includes Meedan (MP0001), UN (MP0002), Multiple-Trans. Ar. Part 1 (LDC2003T18), and Ar. News Trans. Text Part 1 (LDC2004T17) Multiple-Trans. Ar. Part 2 (LDC2005T05). Word alignment is done using GIZA++ (Och & Ney, 2003). For language modeling, the system uses all the monolingual data allowed which are about 850M together with the Arabic side of its training data. The language model is implemented using the IRSTLM toolkit (Federico et al., 2008). Training and decoding were conducted using the Moses phrase-based SMT system (Koehn et al., 2007). The system uses the Penn Arabic Treebank (TB) tokenization scheme to preprocess the Arabic data. The decoding weight optimization was done using a set of 510 sentences from MEDAR Evaluation Campaign 1 evaluation test set (Medar Eval1). The participant produces two outputs. In the primary output, the data is denormalized in which the appropriate form of the Alif and Ya is retrieved in context (enriched form) while in the secondary output, the data is normalized in which all Hamzated Alif forms are converted to bare Alif and dotless Ya/Alif Maqsura is converted to dotted Ya (reduced form) (El Kholy & Habash, 2010).

4.4.2.5. MEDAR Baseline 1 (University of Balamand)

The system is developed on the basis of Moses by the University of Balamand. Two versions of the system have been submitted, according to the (parallel) training and development data used, as presented below:

System	Training data	Development data
Baseline 1-1	Medar_Eval1, MP0001, MP0002	LDC2003T18, LDC2004T17, LDC2005T05
Baseline 1-2	Medar_Eval1, MP0001, MP0002	Baseline

Table 6: Training and development data of the MEDAR Baseline 1 system.

4.4.2.6. MEDAR Baseline 2 (IBM/DCU)

The system is developed on the basis of Moses by IBM in partnership with DCU. Four versions of the system have been submitted, according to the monolingual and parallel training data used, as presented below:

System	Monolingual training	Parallel training
Baseline 2-1	All	All
Baseline 2-2	All	All
Baseline 2-3	Baseline	LDC2003T18, LDC2004T17, LDC2005T05
Baseline 2-4	M0001, M0002, M0003, M0004, M0005, M0006, W0027, W0030, W0036, W0042	Medar_Eval1, MP0001, MP0002

Table 7: Training data of the MEDAR Baseline 2 system.

“Baseline 2-1” and “Baseline 2-2” only differ by the maximum length size of the sentences taken into account: 50 for the former, 100 for the latter.

4.5. Evaluation Schedule

The schedule was specified as follows:

July 08, 2010	Training data are sent to participants
July 23, 2010	Evaluation data are sent to participants
July 28, 2010	Deadline for sending back translations
July 30, 2010	Automatic results are sent to participants

Table 8: Schedule of the second MEDAR evaluation campaign.

4.6. Results

4.6.1. Scoring Tools

Automatic scoring is done by ELDA using BLEU, BLEU/NIST and mWER as for the first evaluation campaign.

4.6.2. Automatic Results

System	BLEU[%]	NIST[value]	mWER[%]
Human reference 1	*69.66*	*12.10*	*25.39*
Google Translate	*20.84*	*6.10*	*66.39*
Sakhr	15.15	5.44	65.85
Univ. of Columbia - Primary	12.59	4.75	75.26
Univ. of Columbia - Secondary	8.54	3.93	79.53
ENSIAS	5.56	3.13	85.67
Baseline 1-2	5.27	3.34	78.47
Baseline 1-1	5.01	3.18	77.73
Baseline 2-1	4.32	2.53	92.37
Baseline 2-2	3.82	2.43	94.70
Univ. of Balamand - Primary	3.79	2.91	78.93
Univ. of Balamand - Secondary	3.77	2.79	84.50
Baseline 2-3	2.82	2.44	90.46
Systranet	*2.03*	*2.12*	*96.70*
Baseline 2-4	0.56	0.81	117.35

Table 9: Results of the second MEDAR evaluation campaign.

4.7. Discussion

The results of this second MEDAR evaluation campaign remain stable compared to the first campaign. Although the test data are different, the results of the two online systems allow us to draw this conclusion since their score did not evolve a lot.

These results are quite surprising since more training data were provided to the systems. More, the results of one of the human reference translation ("Human reference 1") compared to the three other reference translations are higher than for the first evaluation, which would mean the test corpus was more easy to translate or that our translator of phase II is more skilled.

In comparing the two MEDAR baseline systems, it appears that the Baseline 1 obtains significant higher scores and this despite the fact that the Baseline 2 did use more training data.

Regarding the difference between Baseline 1-1 (BLEU score = 5.01) and Baseline 1-2 (BLEU score = 5.27), it seems that adding the LDC data to the development of the system decreased the performance. The baseline development data are then probably better adapted.

Performance of the MEDAR baseline 2 is better when using all the training data for both monolingual and parallel data. However, it highly decreases when using W00XX and M00XX

data only. It seems the system got some issue being adapted to those data. Particularly, we believe the monolingual data did not well improve the language model during the training: The system from ENSIAS and the Baseline 2-3 got better results without the addition of monolingual data than Baseline 2-4 for which monolingual data were added. This clearly means that such data is too heterogeneous with respect to the test corpus.

Lower scores of the baseline systems can be explained by the large number of non recognized words (words out of the system vocabulary and absent from its training data). Most of the low quality translations, especially for the Baseline 2-4, are composed of transliterated words unknown to the system: out of vocabulary words are rendered as they are.

Finally, Baseline 2-1 and Baseline 2-2 are distinguished by the maximum length of sentence used: using shorter sentences gave higher results. This is certainly due to the difficulties of building the translation tables using long sentences.

5. Recommendations

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:

1. Increase the size of training data.
2. Incorporate more tools to account for the specific features of Arabic. We have noticed that the preprocessing used by the Columbia system proved to be efficient.
3. Ensure that the scoring metrics are appropriate for assessing Arabic outputs (e.g. BLEU measures some “consistencies” of n-grams, it may not be adapted to an agglutinative language like Arabic). Human evaluation will be conducted to check this issue. Results of such an evaluation (for instance using Fluency or Adequacy criteria) would allow us to compare human and automatic metrics within a “meta-evaluation” (i.e. the evaluation of the metrics).
4. Improve Moses for Arabic in the same way the University of Balamand did for its own system. These could be: reordering words for alignment, syntactic analysis for preprocessing, segmentation and morphological decomposition, word alignment, etc.

6. Further Work

Despite the low performance achieved by several systems based or derived from Moses, MEDAR is happy to offer these packages to the HLT community. These contain the two baseline systems and the following resources:

- Test and masking corpus of the first evaluation campaign and the four reference translations;
- Test and masking corpus of the second evaluation campaign and the four reference translations;
- MEDAR monolingual training data;
- MEDAR parallel training data.

The current systems are baseline and as such require more improvement, tuning, etc. This should be conducted in a coming initiative. Furthermore, by offering such a package to the universities, students may boost activities on MT for English to Arabic and more largely MT considering Arabic as the target language.

7. References

(Och & Ney, 2003). Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

(Federico et al., 2008). Federico, Marcello / Bertoldi, Nicola / Cettolo, Mauro (2008): "IRSTLM: an open source toolkit for handling large scale language models", In INTERSPEECH-2008, 1618-1621.

(Koehn et al., 2007). Moses: Open Source Toolkit for Statistical Machine Translation, Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, ACL 2007

(El Kholy & Habash, 2010). EL KHOLY A. & HABASH N. (2010). Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC), Valletta, Malta.

8. Annex A. DTD and Example of Input Corpus

```
<!ELEMENT SRCSET (DOC* )>
<!ATTLIST SRCSET setid CDATA #REQUIRED >
<!ATTLIST SRCSET srclang CDATA #FIXED "EN">
<!ELEMENT DOC (seg*)>
<!ATTLIST DOC docid CDATA #REQUIRED >
<!ATTLIST DOC genre CDATA #FIXED "text">
<!ELEMENT seg (#PCDATA)>
<!ATTLIST seg id CDATA #REQUIRED>
<!ENTITY lsquo "&#8216;">
```

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE SRCSET SYSTEM "Corpus_Medar.dtd">
<SRCSET setid="corpus_medar_enar" srclang="EN">
  <DOC docid="1" genre="text">
    <seg id="p1.1">
      Sentence to translate 1
    </seg>
    <seg id="p1.2">
      Sentence to translate 2
    </seg>
```

```
  ...
  <seg id="n">
   Sentence to translate n
  </seg>
 </DOC>
 ...
</SRCSET>
```

9. Annex B. DTD and Example of Output Corpus

```
<!ELEMENT     TSTSET (DOC* )>
<!ATTLIST     TSTSET setid CDATA #REQUIRED >
<!ATTLIST     TSTSET srclang CDATA #FIXED "EN">
<!ATTLIST     TSTSET trglang CDATA #FIXED "AR">
<!ELEMENT     DOC (seg*)>
<!ATTLIST     DOC docid CDATA #REQUIRED >
<!ATTLIST DOC genre CDATA #FIXED "text">
<!ATTLIST DOC sysid CDATA #REQUIRED>
<!ELEMENT seg (#PCDATA)>
<!ATTLIST seg id CDATA #REQUIRED>
<!ENTITY lsquo "&#8216;">
```

Example:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE TSTSET SYSTEM "Corpus_Medar_output.dtd">
<TSTSET setid="enar" srclang="EN" trglang="AR">
 <DOC docid="1" genre="text" sysid="TEST_system">
  <seg id="1">
   Translated sentence 1
  </seg>
  <seg id="2">
   Translated sentence 2
  </seg>
  ...
  <seg id="n">
   Translated sentence n
  </seg>
  ...
 </DOC>
</TSTSET>
```