MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable 5.3

Evaluation methodology and results

Author: Olivier Hamon ELDA, Khalid Choukri, ELDA
Contributors: Chafic Mokbel, University of Balamand, Sara Noeman, IBM Egypt

November 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-Egypt,** The Engineering Company for the Development of Computer Systems, Egypt
- **Birzeit University:** Center for Continuing Education, West Bank and Gaza Strip
- **University Mohammed V Souissi:** 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

European Commission

The MEDAR project is supported by the ICT programme

CONTENT

1 Executive summary

This report deals with the evaluation methodology and results of the MEDAR evaluation campaign. 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[1].

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 evaluation campaign 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 evaluation campaign. 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 on MT evaluation for English-to-Arabic direction in terms of technologies and in terms of resources.

2 Objectives of the MEDAR MT evaluation

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, the MEDAR evaluation campaign targets several objectives:

- Developing a framework for the evaluation of English-to-Arabic MT systems;
- Developing a baseline with background from existing open source tools;
- Producing data for MT training;
- Producing data for MT evaluation;
- Evaluating MEDAR MT baseline systems;
- Comparing MEDAR baseline MT systems regarding other MT systems;
- Creating and federating a new community around the MT English-to-Arabic theme;

[1] The two MEDAR baseline systems are available from the project website.

- Making available a package containing the full set of resources and tools from MEDAR.

3 Baseline systems

In MEDAR, two baseline Statistical Machine Translation (SMT) systems have been used. They are developed by the University of Balamand, UOB ("Baseline1") and IBM with the help of DCU ("Baseline 2") on the basis of Moses[2]. 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.

3.1 MEDAR baseline 1 (University of Balamand)

3.1.1 Existing and selected tools

Several open toolkits exist for Machine Translation and in particular for Statistical Machine Translation. For instance one can cite ***Egypt toolkit***[3] or ***MTTK***[4].

In this project UOB has chosen ***Moses***[5] (Koehn et al., 2007) as the machine translation decoder. Moses has been chosen because it is an open source toolkit. It has proven to be of equivalent quality to proprietary state of the art MT systems. It has been successfully used for different languages. In addition, with Moses, integrating explicit linguistic information is possible. Factored translation models have been included in the implementation (Koehn et al., 2007). Finally, another characteristic that supports the choice of Moses is the confusion network decoding. This facilitates the usage of Moses in a speech-to-speech system by accepting at the input of SMT a network of solutions corresponding to the N-best solution produced by a speech recognizer.

Moses toolkit allows the following:

- ***Preprocessing***: Several perl and shell scripts are included in the toolkit that performs tokenization and detokenization of the input/output text and basic preprocessing of the punctuations.
- ***Language modeling***: In order to perform N-Gram statistical language modeling Moses integrates external open source toolkits such as ***SRILM***[6] (Stockle, 2002).
- ***Modeling, Training and Alignment***: This is done using the ***GIZA++*** tool[7] (Och & Ney, 2003) originally developed within the Egypt toolkit. This implements both HMM and fertility-based models 4 and 5. Giza++ also includes the ***mkcls*** tool. mkcls permits to generate words classes.

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

[3] http://www.clsp.jhu.edu/ws99/projects/mt/

[4] http://mi.eng.cam.ac.uk/~wjb31/distrib/mttkv1/

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

[6] http://www-speech.sri.com/projects/srilm/

[7] http://code.google.com/p/giza-pp/

- ***Tuning***: Moses includes tools to tune the SMT models estimated parameters following the ***minimum error rate training*** (Och, 2003).
- ***Decoding***: Decoding is finding the solution of the ***source-channel approach*** in SMT (Brown et al, 1990), i.e. the fundamental equation of SMT (Brown et al., 1993), given the SMT models and their parameters estimated and tuned.

3.1.2 Systems developed and obtained results

The work conducted within Workpackage 5 (WP5) has been split into two phases. In a first phase, a baseline system has been built using Moses. Variants to this system have been studied in the second phase. The baseline system has been developed for bidirectional English-to-Arabic and Arabic-to-English translations. The variants were only for English-to-Arabic translation.

3.1.2.1 Baseline system

A set of scripts have been developed for an easy install and use of the Moses system. The script verifies if a tool package (SRILM, GIZA++, Moses) is already installed. If a tool is not installed, it will install it. The only modification performed in the Moses set of tools is in the preprocessing tool taking into account some specificities of Arabic. Actually, the existing preprocessing tools tokenize the text (i.e. arrange the spaces in the sentence), filter out the long sentences and lowercase all the characters. For the Arabic language, some punctuations are different and there is no upper case. This has been taken into account.

Once installed, the baseline may be used either to build a SMT model and evaluate it or only to use it. The first script permits to build the SMT models and evaluate it. The block diagram is shown in

Figure 1. One may distinguish the three phases: training, tuning and, decoding and evaluation. It is worth noting that the parallel English-Arabic database is split into three parts: training set (the largest one), development set to tune the SMT model parameters and, testing set to decode and evaluate the resulting translated text compared to the parallel text in the target language. Finally, the evaluation tool permits to compute BLEU and NIST scores.

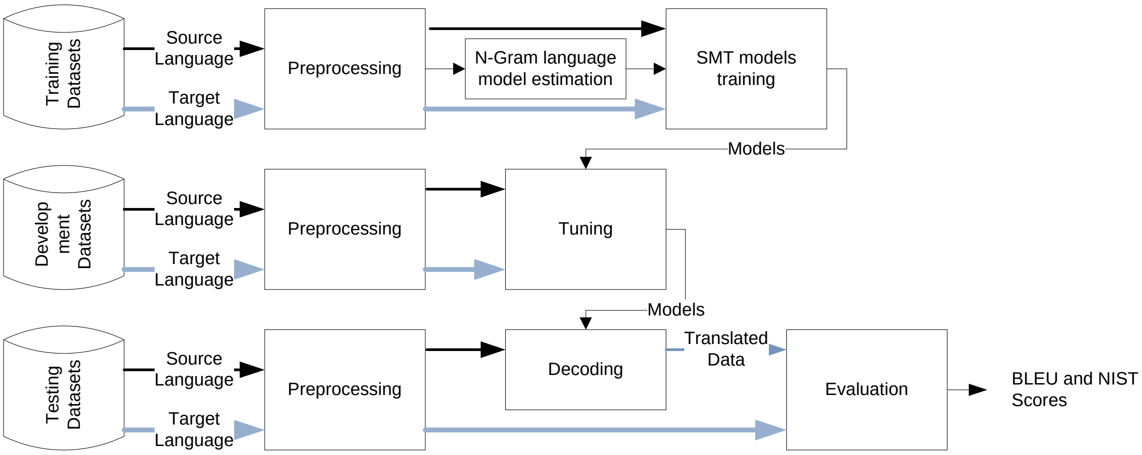

Figure 1: The Baseline system based on Moses.

Besides the complete training-decoding-evaluating system, another script has also been developed to perform only translation. This script installs only a reduced package with the necessary Moses tools to perform translation and includes the pre-trained and pre-tuned SMT models.

3.1.2.2 Hybrid systems

In order to improve the performance of the system two aspects have been explored.

3.1.2.2.1 Morphological information

UOB has developed a limited morphological analysis system (Ghaoui et al., 2005). It is a finite state machine as shown in the Figure 2. The different elements of this stemmer are:

- AL : ‘ال’, ‘وال’, ‘فال’, ‘بال’, ‘كال’, ‘لل’
- GEN : ‘و’, ‘ف’, ‘ب’, ‘ك’, ‘ل’, ‘س’
- PLUR : ‘ان’, ‘ين’, ‘ون’
- POS ‘تموني’, ‘تمونا’, ‘تهما’, ‘وهما’, ‘تيني’, ‘تينا’, ‘تموني’, ‘تلكَ’, ‘تلكِ’, ‘تني’, ‘تنا’, ‘تكم’, ‘تكن’, ‘تهم’, ‘تهن’, ‘تها’, ‘وكَ’, ‘وكِ’, ‘وكم’, ‘وكن’, ‘وها’, ‘وهم’, ‘وهن’, ‘وني’, ‘ونا’, ‘هما’, ‘تً’, ‘تِ’, ‘ك’, ‘ك’, ‘كم’, ‘كن’, ‘ني’, ‘ها’, ‘هم’, ‘هن’, ‘وا’, ‘وك’, ‘وه’, ‘تك’, ‘ته’, ‘تم’, ‘نا’, ‘تَ’, ‘ك’, ‘ه’, ‘ي’, ‘ت’

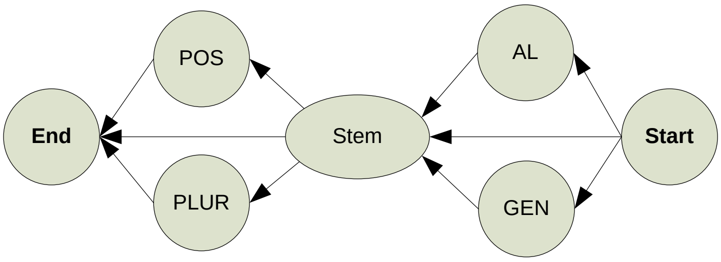

Figure 2: State machine simple stemmer.

In the first variant developed this analysis is applied on the text and the most frequent prefixes and suffixes are separated from the words and considered as independent words. After translation, prefixes and suffixes are rearranged with their corresponding words.

3.1.2.2.2 Synonyms

After the first experiments we have identified synonyms as a major source of errors committed by the baseline system. The system was often translating to the most frequent of the synonyms. The idea of this variant is to add in the training different suffixes to the English words depending on the translated synonyms. During the decoding the words that have equivalent synonyms in the target language will be appended with the different synonyms and the phrase translation with the highest score is kept.

3.1.2.3 Experimental results

The baseline system and the two variants have been originally experimented on the following parallel corpora:

- Arabic News Translation Text Part 1 by LDC (ref. LDC2004T17 and ISBN 1-58563-307-0). It has a total of about 440,000 Arabic words collected from AFP, Xinhua and Annahar.

- Arabic English Newswire Translation Collection by LDC (ref. LDC2009T22 and ISBN 1-58563-521-9). It has a total of about 550,000 Arabic words collected from AFP, Annahar and Assabah.

The experiments results yielded to BLEU scores of 2.22 for the baseline system, 2.7 for the variant with morphological analysis and 2.88 when both synonyms and, prefixes and suffixes separation are considered. The synonyms variant when experimented in the MEDAR evaluation campaign has drastically degraded the performances. This problem is still under investigation.

3.2 MEDAR baseline 2 (IBM Egypt / Dublin City University)

3.2.1 Moses toolkit

The IBM/DCU baseline system is also based on Moses that has been described in the previous section. Basic technical details are given herein and apply to both systems.

3.2.2 Basic components – used toolkits

The SRILM Toolkit has been used as language model and GIZA ++ Toolkit as translation model for word alignments and Heuristics to build the phrase table. The decoder used is the Stack decoding algorithm.
It requires:

- Phrase Table: Phrase Translation table
- Moses.ini : The configuration file for the decoder
- Language Model File
- Running decoding: *echo 'this is a small house' | moses -f moses.ini > out*

3.2.3 Alignment toolkit

GIZA++ is used with parameters describing input files:

- c = (training corpus file name)
- s = (source vocabulary file name)
- t = (target vocabulary file name)
- Other parameters that can modify the models, EM-algorithm, smoothing parameters, etc...

However, Moses training script train-factored-phrase-model.perl calls GIZA++ internally for the training of statistical translation models. GIZA++ is an extension of the program GIZA (part of the SMT toolkit EGYPT[8]) which was developed by the Statistical Machine Translation team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns Hopkins University (CLSP/JHU). GIZA++ includes a lot of additional features. The extensions of GIZA++ were designed and written by Franz Josef Och[9]. The program includes the following extensions to GIZA:

- Model 4;
- Model 5;

[8] http://www.clsp.jhu.edu/ws99/projects/mt/toolkit/

[9] http://www.isi.edu/~och

- Alignment models depending on word classes (software for producing word classes is included in the package);
- Implements the HMM alignment model: Baum-Welch training, Forward-Backward algorithm, empty word, dependency on word classes, transfer to fertility models;
- Includes a variant of Model 3 and Model 4 which allows the training of the parameter p_0;
- Various smoothing techniques for fertility, distortion/alignment parameters;
- Significantly more efficient training of the fertility models;
- Correct implementation of pegging as described in Brown et al. 1993, a series of heuristics in order to make pegging sufficiently efficient.

3.2.4 Language model toolkit

SRILM is an Open Source package for building LM using pre-processed text with the main commands:

1. **ngram-count:** takes a text file as input, generates an intermediate count file and an n-gram language model (it can also use count file as input)
2. **ngram-merge:** can merge large count files for parallel work

The ngram-count script uses the following options:

- *text textfile*: Input text file used to generate the LM textfile should contain one sentence unit per line;
- *order n*: Set the maximal order (length) of N-grams to count, the default order is 3. (i.e. 3-gram model);
- *lm lmfile*: Estimate a backoff N-gram model from the total counts, and write it to lmfile;
- *write-binary-lm*: Generates a binary LM instead of text (save storage);
- *unk*: Build an "open vocabulary" LM, i.e. the unknown-word token is used as a regular word, the default is to remove the unknown word;
- *vocab file*: Read a vocabulary from file. Subsequently, out-of-vocabulary words in both counts and text are replaced with the unknown-word token. If this option is not specified all words found (in corpus) are implicitly added to the vocabulary.

3.2.5 Moses training

Given a pair of parallel corpus files, Moses can generate a standard phrase model using the main script. A factored model can also be generated by adding factors (POS tags, lemmas, etc) beside each word in the corpus. The following steps have been run:

1. Prepare data
2. Run GIZA++
3. Align words
4. Get lexical translation table
5. Extract phrases
6. Score phrases
7. Build lexicalized reordering model
8. Build generation models
9. Create configuration file

The basic parameters are:

- max-phrase-length (default 7 words): Shorter phrases may be needed if phrase table size is large;
- giza-option: Additional options for GIZA training;
- translation-factors (for Factored-Models) Create one or more translation tables between a subset of the factors.

We used the standard command *moses -f moses.ini -i in.file* > *out.file*, using:

- moses.ini : moses config file
- in.file : input file (source language)
- out.file : output file (target language)

More advanced options are:

- -t : trace, reveals which phrase translations were used;
- -v : verbose, displays additional run time information;
- These are mainly used for debugging and optimization.

3.3 Preprocessing toolkits

3.3.1 AMIRA-1.0

The ArabicSVMTools package (Tools for processing Arabic text from raw text to Base Phrase Chunks) has separate modules for the processing of Arabic script. It takes a regular transliterated Arabic text file and produces it tokenized, part of speech tagged and base phrase chunked.

The system was developed, trained and tested on the Arabic Penn TreeBank ATB 1 v3.0, ATB 2 v2.0 and ATB 3 v2.0. You can find detailed information about the Arabic TreeBank corpus in the LDC release. But briefly, the corpus is from AFP and it is newswire covering domains such as politics, and sports. The contents of the package are:

./bin	# The relevant scripts
./tokmodels	# Trained models for Tokenization
./lemmodels	# Trained for changing the t to p for singular feminine nouns when followed by possessive pronouns
./posmodels	# Trained models for POS tagging
./bpmodels	# Trained models for BP chunking
./example	# Sample document in buckwalter's Transliterated scheme and its tokenization feminine Lemmatization, POS tagged version and BPchunked form. The file names are indicative of the contents.

Yamcha[10] is mandatory to use AMIRA.

3.3.2 OpenNLP-toolkit

OpenNLP-toolkit is a set of java-based NLP tools which perform sentence detection, tokenization, pos-tagging, chunking and parsing, named-entity detection, and co-reference.
These tools can be integrated with other software to assist in the processing of text.

[10] http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/

4 Guidelines for the production of evaluation data

In order to produce reference translations of good quality, MEDAR defined guidelines for two main steps of the evaluation campaign: the human translation of test data that will serve as references and the validation of those translations.

4.1 Translation guidelines

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 represents the target against which to compare the MT systems outputs.

4.1.1 Translation team

A single translation team is used to translate all of the source language data. This team is composed of:

- Several bilingual translators, native speakers of the target language of the data (Arabic).
- A bilingual but Arabic 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.

Translations are systematically finalized and checked by an Arabic 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.

4.1.2 Test material

Data are monolingual texts coming from a specific domain and have an average length of twenty words per sentence. They may come from websites 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. Actually, the use of these websites is strictly forbidden. The translated file is rendered in XML format, UTF-8 encoded, so as to preserve the original structure.

4.1.3 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.

4.2 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 described herein.

4.2.1 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 deadline 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.

4.2.2 Formal validation

The first validation process consists of an automatic validation that 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 schema given 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.

4.2.3 Content 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 delivered 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 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 translation receives the “benefit of the doubt”. Only clear errors should be indicated.

When several translations are produced for the 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.

4.2.4 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.

4.2.5 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.

5 MEDAR language resources

Three types of data have been produced and distributed within the MEDAR evaluation campaign: monolingual training data, parallel training data and evaluation data.

5.1 File format and DTD

Four types of resources are considered within MEDAR: monolingual corpus, parallel corpus, input for the evaluation and output of the evaluated systems. Each corpus is encoded in XML and UTF-8 and contains documents identified with a *docid* attribute. Documents are segmented in sentences. Each sentence within a document is tagged and identified with an *id* attribute. The format specifications of the corresponding DTD and examples are given in Annexes for each type of corpus.

The output files must preserve the original structure and a *sysid* attribute is added to the DOC tag.

5.2 Training data

Training data are used to train the MT systems prior to the evaluation campaign. They are of two types: monolingual corpora and parallel corpora.

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.

5.2.1 Preparation

5.2.1.1 Monolingual data

Three sources have been used to produce the MEDAR monolingual corpus. ELRA and LDC 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, composed of articles from newspapers, have 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[11].
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 result of the project.

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

[11] http://download.wikipedia.org/backup-index.html

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 Evaluation only
Al-Hayat	ELRA-W0030	38M	MEDAR Evaluation Only
LMD	ELRA-W0036	475K	MEDAR Evaluation Only
NEMLAR	ELRA-W0042	494K	MEDAR Evaluation Only
Arabic Gigaword 4th Ed.	LDC2009T30	2GB	MEDAR Evaluation only

Table 2: Monolingual data used for training.

5.2.1.2 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 dry-run. It consisted of the test corpus and the four “reference” translations, formatted into four parallel corpora of 10K words (see below).

Two parallel corpora have been selected from already existing data: Meedan translation memory composed of news articles, and UN corpus originally available from http://www.uncorpora.org. The latter is composed of collections from the United Nations General Assembly Resolutions.

Crawling and formatting have been made using our own scripts since the task was quite simple. One could have used more powerful tools such as *bitextor*[12] that is an automatic bitex generator using data from the Internet, or *combine*[13] that may be used as a focused crawler.

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.

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

[12] http://bitextor.sourceforge.net

[13] http;//combine.it.lth.se

Name	Id	Size [words]	Availability
MEDAR Dry-run	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 Evaluation only
Ar. News Trans. Text Part 1	LDC2004T17	441K	MEDAR Evaluation only
Multiple-Trans. Ar. Part 2	LDC2005T05	15K	MEDAR Evaluation only

Table 3: Parallel data used for training.

5.3 Evaluation data

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.

5.3.1 Material

Input data are English texts coming from a specific domain, the climate change.

5.3.2 Preparation

The overall evaluation data have been built as follows:

1. 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, 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.).

The test corpus has 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, cf. section 4.

For the dry-run, the evaluation data 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.

For the evaluation campaign, the evaluation data 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. We decided to reduce the masking corpus after the dry-run experience since participants had a short delay to produce the translation and because the evaluation data was already large enough.

6 Scoring tools

We evaluated the systems using both automatic and human evaluations.

6.1 Automatic evaluation

Automatic scoring was 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 weights 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).

6.2 Human evaluation

For all the systems, each sentence is evaluated in relation to adequacy and fluency measures. For the evaluation of adequacy, the target sentence is compared to a reference sentence. For the evaluation of fluency, only the syntactical quality of the translation is evaluated. The evaluators grade all the sentences firstly according to fluency, and then according to adequacy, so that both types of measures are done independently, but making sure that each evaluator does both for a certain number of sentences.

For the evaluation of fluency, evaluators have to answer the question: “Is the text written in good Arabic?”. A five-point scale is provided where only extreme marks are explicitly defined, ranging from “Perfect Arabic” to “Non understandable Arabic”. For the evaluation of adequacy, evaluators have to answer the question: “How much of the meaning expressed in the reference translation is also expressed in the target translation?”. A five-point scale is also provided to the evaluators, where, once again, only extreme cases are explicitly defined, going from “All the meaning” to “Nothing in common”.

Two evaluations are carried out per sentence, they are done by two different evaluators, and sentences are distributed to evaluators randomly, because evaluators should not build a storyline and preserve information between two adjoining segments.
Evaluators are native speakers of Arabic educated up to university level.

7 Evaluation

7.1 Dry-run

7.1.1 Training

There was no training or development phase planned for the dry-run, therefore no data was provided to participants. The two MEDAR baseline systems have not been specifically trained and a very basic data set has been used, corresponding to 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.

7.1.2 Participating systems

The two baseline SMT systems have been used. They are developed by the University of Balamand ("Baseline1") and IBM/DCU ("Baseline 2") on the basis of Moses. Furthermore, the evaluation campaign was open to external participants and participants from the MEDAR consortium, and so was the dry-run. Therefore, a promotion of the campaign has been made through several procedures: mailing lists, networking, personal contacts, conferences, etc. Four participants replied and five submissions have been made. The modest participation may be explained by the short delay between the start of the campaign and the scoring. However, it also may be due to the lack of existing English-to-Arabic systems in the field. For this dry-run, five submissions have been received, anonymized and renamed as "System A" to "System E".
Finally, for comparison purposes, two online systems have been used: Google Translate[14] and Systranet[15]. Their results must be considered carefully since they are not really participating systems.

7.2 Schedule

The schedule of the dry-run 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 4: Schedule of the MEDAR dry-run.

7.3 Results

7.3.1 Automatic (anonymized) results

Results have been automatically computed against four references. To compare to what a human translator can produce and to put into perspective the results of the

[14] http://translate.google.fr/?hl=fr&tab=wT#

[15] http://www.systran.fr/

automatic systems, the results of one (arbitrary) reference translation (*Human reference 1*) is presented below, comparing it against the three other reference translations (as if the translator 1 is considered as a "perfect" MT system). Results are shown in Table 5.

System	**BLEU [%]**	**NIST [values]**	**mWER [%]**
Human reference 1	*56.3*	*11.0*	*28*
Google Translate	*20.3*	*7.0*	*68*
System A	16.6	6.3	67
System B	11.7	4.8	73
System C	11.2	5.0	76
System E	5.9	3.5	78
System D	5.7	3.9	79
Baseline 1	5.1	3.7	81
Baseline 2	4.5	3.6	86
Systranet	*2.1*	*2.3*	*107*

Table 5: Anonymized results of the MEDAR dry-run.

7.3.2 Human evaluation results

7.3.2.1 Setup

For the human evaluation, 12 "systems" have been evaluated: the 10 systems presented in Table 5, plus two systems for which remaining English words in translation have been replaced by several "*" characters. This should allow us to observe the influence of the non translated words on judges. These two systems are named as "System D_bis" (corresponding to the "System D") and "Baseline 1_bis" (corresponding to the "Baseline 1").

Therefore, 6,120 sentences were evaluated twice and randomly distributed among 50 different judges. It represents around 245 sentences per judge. Unfortunately, only 11 judges proceeded to the evaluation against our expectations. It represents 1,548 sentences evaluated, being around 129 sentences per system.

7.3.2.2 Results

Human evaluation results of the dry-run are shown in Table 6.

System	Fluency [1-5]	Adequacy [1-5]
Human reference 1	*4.18*	*4.30*
Google Translate	*3.46*	*3.64*
System A	3.16	3.36
System B	2.55	2.88
System C	2.82	3.19
System E	2.06	2.15
System D	2.03	2.16
Baseline 1	1.98	2.06
Baseline 2	1.70	1.61
Systranet	*1.96*	*2.18*
System D_bis	1.87	2.04
Baseline 1_bis	1.80	2.15

Table 6: Human evaluation results of the MEDAR dry-run.

7.3.2.3 Analysis

The modest number of participants obviously limits the interest of this evaluation. This may be due to the period of the evaluation (summer break), the late evaluation regarding the period we contacted judges and a lack in motivation (certain judges did start the judgements and stop after they notice the difficulty or that it is not a pleasant task). We took those remarks into consideration for the MEDAR evaluation campaign.

However, as a dry-run, this human evaluation confirmed that the protocol was correct and that all the tools (interface, preparation scripts, metrics) were working.

Results show a 98% correlation between BLEU and the adequacy scores and a 96% correlation between BLEU and the fluency score.

7.3.3 Discussion

For this dry-run 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. The evaluation has been useful 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 human reference translation (“Human reference 1”) scores lower than we expected, and the four translations are similar. 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, one could argue that BLEU or any current automatic metric may be not adapted to process Arabic data particularly due to the agglutination features of Arabic. However, results seem to provide a very good correlation with human metrics, much higher than for other languages (Callison-Burch et al., 2010).

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

7.4 Evaluation campaign

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

7.5 System training

7.5.1 Training conditions

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

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

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.

Unfortunately, we received no participation to the Unconstrained Condition. Therefore, we consider only the Constrained Condition in the results shown below. All the participants used the training data provided by MEDAR, except for one rule-based that obviously used its own data.

7.5.2 Participating systems

7.5.2.1 Overview

As for the dry-run, 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. Four submissions from the two MEDAR baseline systems have been made.

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 in 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 vis-à-vis statistical MT systems.

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

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. The descriptions below are provided by the participants in the campaign.

7.5.2.2 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.

7.5.2.3 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 MT system is the Morphological Analyzer. The analyzer is based on an Arabic lexicon that contains valid stems along with their part of speech (POS), root and pattern, applicable prefixes and suffixes, morphological features (e.g. gender, number, person), syntactic features (e.g. transitivity, agreement, pre-terminals), and semantic features (e.g. senses, taxonomies, attributes). For each Arabic 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 Analyzer uses other resources: a statistical POS tagger and Named-Entity recognizer as well as a database of common spelling mistakes and an Arabic Language model for text verification and name detection. The output of the morphological analyzer is used in subsequent steps of the Sakhr MT process.

The second step in the Sakhr MT process is automatic diacritization. In addition to stem diacritization, the Sakhr automatic diacritizer assigns case ending diacritics at the end of verbs and nouns. The verb cases are the indicative, subjunctive, and jussive. For the nouns, the cases are nominative, accusative, and genitive, which could be applied with or without nunation, depending on the definiteness of the noun. The case ending diacritics are determined using rules that depend on adjacency relations with function words like prepositions, articles, demonstrative articles, pronouns, relative pronouns, etc. They also determine case endings for different syntactic structures like noun-noun, noun-adjective, and verb-subject-object relations, with the help of agreement conditions and a selection restriction database. Expressions (e.g., proper nouns, idioms, adverbs, and collocations) are saved in their fully diacritized form whenever possible, to enhance diacritization accuracy. The accuracy of the diacritizer measured on a validation set is 97% for stem diacritization, and 91% for full diacritization.

The final phase in the Sakhr system is machine translation itself. This process uses the information from the components described above to disambiguate the Arabic words, and assign feature values to them. This input is used, together with Arabic grammar rules to produce a full parse of the source sentence. Transfer rules, and an Arabic-to-English lexicon are then used to transform the Arabic parse tree to English. 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 English expressions.

7.5.2.4 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.

7.5.2.5 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 produced 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).

7.5.2.6 Setup of the MEDAR baseline 1 (University of Balamand)

The system is developed on the basis of Moses by the University of Balamand. One version of the system has been submitted using (parallel) training and development data presented in Table 7:

System	Training data	Development data
Baseline 1-1	All corpora	Baseline

Table 7: Training and development data of the MEDAR Baseline 1 system.

7.5.2.7 Setup of the MEDAR baseline 2 (IBM Egypt / Dublin City University)

The system is developed on the basis of Moses by IBM in partnership with DCU. Three versions of the system have been submitted, according to the monolingual and parallel training data used, as presented in Table 8.

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

Table 8: 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.

7.5.3 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 9: Schedule of the MEDAR evaluation campaign.

7.6 Analysis of the parallel training data

We decided to split the parallel training data in two parts due to the difference in usage rights: LDC and MEDAR.

The LDC training data refer to LDC2003T18 (Multiple-Translation Arabic Part 1), LDC2005T05 (Multiple-Translation Arabic Part 2) and LDC2004T17 (Arabic News Translation Text Part 1) resources that correspond to newswires from two sources of Arabic data (Xinhua News Service and AFP News Service, and An Nahar for LDC2004T17 only).

The MEDAR parallel training data refer to Medar_Eval1 (data from the dry-run and from the climate change domain), MP0001 (Meedan translation memory containing news data) and MP0002 (United Nations data).

Main data are close to news or diplomatic domain, but are quite heterogeneous. The different corpora contain a lot of proper names (many are different from those of the test corpus).

We conducted a comparison between the training copora and the test corpus. We particularly focused on the vocabulary used and the size of the lexicon. To do so, we simply computed the number of different English words for both LDC and MEDAR parallel corpora. Results are shown in Table 10: Statistics on training and test corpora.

Corpus	#English Lexicon	Mean of word occurence	Median
LDC	27,276	28.5	2
MEDAR	28,797	91.3	3
LDC+MEDAR	41,789	81.6	2
Test	2,444	3.7	1

Table 10: Statistics on training and test corpora.

Both LDC and MEDAR training corpora are quite similar in terms of distinct number of words. Half the words are in the two corpora. Comparing the means to the medians, we can see that most of the words are not present a lot in the corpora: a few words a far more frequent, such as non-content words ('the', 'a', etc.). Means of words number indicate that LDC parallel corpus is more heterogeneous than the MEDAR one. 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. In the same way, difference in number of words that appears 2, 3, or less than 100 times remains stable between the two corpora. Finally, there are more words that are very frequent in the MEDAR corpus than in the LDC corpus.

We then compared the training corpora to the test corpus. Table 11 shows the out-of-vocabulary of the test corpus regarding both LDC and MEDAR corpora as well as the overall parallel training corpus.

Corpus	# different words	# different unknown test words	# words	# unknown words
Test	2,444	-	8961	-
LDC	27,276	384 (16%)	778,682	604 (7%)
MEDAR	28,797	250 (10%)	2,630,330	388 (4%)
LDC+MEDAR	41,789	194 (8%)	3, 409,012	306 (3%)

Table 11: Out-of-vocabulary of the test corpus.

A substantial part of the lexicon is unknown to the MT systems when translating the test corpus. When training the MT system using the LDC training corpus, around 16% of the tests corpus lexicon is unknown, that is quite important. Proportions are still important using the MEDAR training corpus (10%) or the overall training corpus (8%). However, unknown words are not less frequent, since the proportion of unknown words is lower than the proportion of different unknown words. For instance, 3% of the test corpus words are unknown using the overall training corpus. Therefore, mWER can not reach less than 3% for every system that has been trained with this corpus. This is worse for BLEU score, since it uses n-grams and that a maximum of 3% of the n-grams (*n* being hereafter equal to 4) may be not found. This is then an argument for using better parallel training corpora, namely one that fits properly to the test corpus.

7.7 Results

7.7.1 Human evaluation results

7.7.1.1 Setup

For the human evaluation, 10 "systems" have been evaluated: 4 primary MT systems, 3 baseline MT systems, two online systems and one human reference. Each system contained 396 sentences. Therefore, 3960 sentences were evaluated twice (giving a total of 7,920 sentences) and randomly distributed among 50 different judges. It represents around 158 sentences per judge.

7.7.1.2 Inter-judge *n*-agreement

To test the agreement among judges, we compute the 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 12: Inter-judge n-agreement [0-1] of the MEDAR evaluation campaign.

Results are similar to previous experiments: for close to 40% of the evaluated sentences, judges give similar scores, that 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 has been done in correct conditions.

Again, as in previous experiments, adequacy *n*-agreements are lower than fluency *n*-agreements, meaning that judgements are more complex for adequacy than for fluency.

7.7.1.3 Results

Human evaluation results are shown in Table 13.

7.7.1.4 Analysis

The human evaluation shows a clear hierarchy among the translations. Human translation obtains high results, but not higher than expected. This is similar to other campaigns in the MT field: translations are not perfect, and human judgment can differ from the conception of a sentence meaning. Moreover, comparing two human translations corresponds to the disagreement between two translators.

Google Translate, Sakhr and University of Columbia results are, in this order, all above 3 points in both fluency and adequacy. Their outputs provide almost understandable translations.

Systranet, University of Balamand and the three MEDAR baseline systems results are under average, providing translations difficult to understand. ENSIAS results are lower.

System	**Adequacy [1-5]**	**Fluency [1-5]**
Human reference 1	*4.34±0.07*	*4.11*±0.08
Google Translate	*3.45*±0.10	*3.49*±0.08
Sakhr	3.27±0.09	3.26±0.08
Univ. of Columbia - Primary	3.07±0.10	3.30±0.09
Baseline 1-1 (MEDAR+LDC train / Baseline dev)	2.34±0.09	2.12±0.09
Systranet	*2.23*±0.08	*2.05*±0.08
Univ. of Balamand - Primary	2.17±0.09	1.92±0.08
Baseline 2-2 (LDC parallel & baseline mono)	2.16±0.10	1.83±0.08
Baseline 2-3 (MEDAR parallel & mono)	2.03±0.09	1.74±0.08
ENSIAS	1.77±0.07	1.41±0.05

Table 13: Human evaluation results of the MEDAR evaluation campaign ranked according to adequacy scores.

Regarding the baseline systems, results are higher when using the overall parallel training corpus. However, results are surprisingly higher using LDC parallel training corpus than the MEDAR one. This is surprising since the unknown words proportion is higher for the LDC parallel training corpus. One explanation may be the usage of the monolingual corpus that could have deteriorated the quality of the translations.

Similarly to other evaluations, fluency results are lower than adequacy ones. This is mainly due to the presence (adequacy evaluation) or not (fluency evaluation) of a reference translation for comparison. Without any reference point, judges tend to be more strict, in case of doubt.

Looking at the human judgements in details, we identified five general problems the MT systems may have to address (several examples are also given in Annex E):

- Missing lexicon entries: out-of-vocabulary words are either kept in English (i.e. latin encoded) or transliterated. Obviously, English words affect the quality perceived by human judges. Transliterated words are either hardly understandable by human judges – because of a specific vocabulary not so close to their knowledge – or contain one or several latin characters that causes definitly the incomprehension of the words, and generally what is said. It also appears that some good transliterations are not well scored by the human judges due to either a lack of knowledge or another existing word for the translation in Arabic.

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

Table 14: Example of unknown words (fluency=1; adequacy=1).

Source	The next step will be promoting the adoption of the principles of integrated natural resource management and the ecosystem approach (land, water and species) in all development project and initiatives along the Zarqa River Basin.
Reference	سوف تكون الخطوة التالية هي تشجيع اعتماد مبادئ الإدارة المتكاملة للموارد الطبيعية وتوجه النظام البيئي (الأرض، والماء والأصناف الأحيائية) في كل مشروعات ومبادرات التنمية على طول حوض نهر الزرقاء.
Translation	سيروّج الخطوة تالي كنت التبن من المبادئ من يضمن [نتثرل رسورس] إدارة والنظام بيئيّ مقاربة (أرض, ماء ونوع) في كلّ تطوير مشروع ومبادرات على طول [زرقا] نهر حوض.

Table 15: Example of transliterated words (fluency=2; adequacy=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.

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

Table 16: Example of issue with compound words (fluency=5; adequacy=1): "clear" is translated as a word instead of "clear distance".

- Complex sentences (comprising coordinated structures, subordinated structures or sentences, etc.) translation: there are syntactic issues when translating complex sentences. Complex sentences may not be identified as such or segments may not be split correctly. This implies that the translation is not focused on the correct meaning. This is particularly so when sentences are long. Generally speaking, the longer the sentence, the more chance is there to have syntactic issues due to the weak identification of the sentence construction. This is the case for our baseline systems, but better systems such as Sakhr or Google Translate are also concerned.

Source	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:
Reference	إنهم يحسبون انبعاثات الطيران العالمي مستقبَلاً من ثاني أكسيد الكربون وأكاسيد النِتْرِيكْ بفِعْل الحركة الجوية حسب أربعة سيناريوهات (المجلس الحكومي الدولي للتغير المناخي/ التقرير الخاص عن سيناريوهات الانبعاثات):
Translation	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 17: Example of issue with a complex sentence (fluency=3; adequacy=1): not all the parts of the sentence are correctly translated; several dependent clauses are hard to split and proper names are missing.

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

Source	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.
Reference	pH لقد اكتشفوا أن ماء البحر يبدل في كِمْياء الجُوِيثايت، حيث توجِد مستويات المنخفضة في المياه شحنة موجبة على سطح رواسب الجُوِيثايت، مما يجعلها جاذبة للزرنيخ سالب الشحنة.
Translation	, حيث منخفضة eتهتeمِستري ل غْeecرس ب er لَتeوَتَeو اكتشفت أن س نتس , eدِم سeتهتeةح مستويات المياه في تهيئة إيجابية تغيير في السطحية ل غْ eنcمما يجعل ها جاذبية ب شكل سلبي متهمةَرس.

Table 18: Example of bad lemmatization (fluency=1; adequacy=1): words have been cut and then not translated but simply transliterated.

- Named entities translation: a lot of named entities are not translated or not well transliterated. This is above all due to some lack in lexicon in the training data. It causes a strong decrease of the fluency (when the translation is poor, missing named entities doesn't help to rebuild correctly the sentence) and less frequently also the adequacy. Indeed, missing named entities does not imply the meaning is hard to found (e.g. we can understand that somebody did something without knowing who did it: in a certain way, this is not crucial to understand the translation).

Source	Hemlock Semiconductor just started building a polysilicon plant in Tennessee.
Reference	وشرعت هِيمْلُوك لأشباه الموصلات للتوّ في بناء مصنع للبُولِي سِيلِيكون في تِينِيسِي.
Translation	hemlock semiconductor مسلح لبناء polysilicon اليرموك في.

Table 19: Example of named entities not translated and wrong word order (fluency=1; adequacy=1).

In particular for the baseline systems, we observed typical errors according to the level of fluency score. When too many words are not translated, especially named entities, the fluency score is often put at its lower level. A fluency score of 2 (second lower level) is generally linked to a wrong generation and rebuild of the sentence in the target language.

Here, the language model shows its limits. Moreover, the Arabic morphology is not well respected: many suffixes or prefixes are not agglutinated properly as it should be. Fluency scores of 3 (mean score) and 4 (close to be a very good translation) correspond to different levels of problems regarding the semantic rendered in syntax or, more often, source sentences in English that are complicated with over three or four connected clauses with, for instance, number and gender badly rendered in the Arabic syntax:

Source	Over half of those live near the coast, making them directly vulnerable to sea-level rise.
Reference	أكثر من نصف هؤلاء يسكنون بجوار الساحل، مما يجعلهم معرَّضين مباشرة لارتفاع منسوب سطح البحر.
Translation	أكثر من نصف هم يعيشون ب القرب من الساحل , مما يجعل ها مباشرة عرضة ل ارتفاع منسوب مياه البحر .

Table 20: Example of named entities not translated and wrong word order (fluency=3; adequacy=4).

In the same way, adequacy scores are affected by typical errors, starting by similar ones to those of the fluency. Because of the pretty low translation quality level, a not fluent translation affects the understanding of the meaning. As already said, there is an important amount of out-of-vocabulary words. Furthermore, numbers in numerical characters causes some issues to the MT systems in wrongly translating the corresponding term (for instance “2 actions” translated into “2 years”): the translation model has been perturbed by a mistranslation in the training data. Finally, segmentation in the English source may be wrong due to a lack of lemmatisation.

There is also a large number of sentences that are fluently correct (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 test corpus has been complex to handle due to its quite specialized domain. It is the case for the MT systems as well as for the judge that may happen to have a lack of knowledge of a certain lexicon.

7.7.2 Automatic evaluation results

7.7.2.1 Setup

After the participants sent back the translation of their systems, files’ format was checked and corrected (in case of little mistakes such as a missing tag) or sent back to participant when the format contained too much garbage. Files were prepared so as to be evaluated quickly with the same evaluation scripts using an evaluation platform.

7.7.2.2 Results

Automatic results are shown in Table 21.

System	BLEU [%]	NIST [value]	mWER [%]
Human reference 1	*69.7*	*12.1*	*25*
Google Translate	*20.8*	*6.1*	*66*
Sakhr	15.2	5.4	66
Univ. of Columbia - Primary	12.6	4.8	75
Univ. of Columbia - Secondary	8.5	3.9	79
Baseline 2-3 (MEDAR parallel & mono)	6.5	3.5	88
Baseline 2-2 (LDC parallel & baseline mono)	6.3	3.5	87
Baseline 2-1 (MEDAR+LDC parallel & Mono	6.1	3.4	89
Baseline 1-1 (MEDAR+LDC train / Baseline dev)	6.1	3.7	76
ENSIAS	5.6	3.1	86
Univ. of Balamand - Primary	3.8	2.9	79
Univ. of Balamand - Secondary	3.8	2.8	85
Systranet	*2.0*	*2.1*	*97*

Table 21: Results of the MEDAR evaluation campaign.

7.7.2.3 Analysis

Ranking results are quite different to those of the human evaluation in the second part of the table. The order of the baseline systems is reversed. Although this is a bit surprising, translations are very close and these differences are not significant enough to draw any conclusion.

Systranet results are explained by the well-known bias that occurs when using n-grams-oriented metrics on rule-based MT systems.
University of Balamand results are also surprising. Here, we assume that judges have been influenced by number of untranslated English words in the Arabic translation.

7.7.3 Meta-evaluation

The human judgements allow us to evaluate the efficiency of the automatic metrics. We then compare the automatic scores to the human ones in order to test their correlation. Both fluency and adequacy scores have been tested against BLEU, NIST and mWER automatic metrics. A comparison of both automatic and human scores is presented in Figure 3.

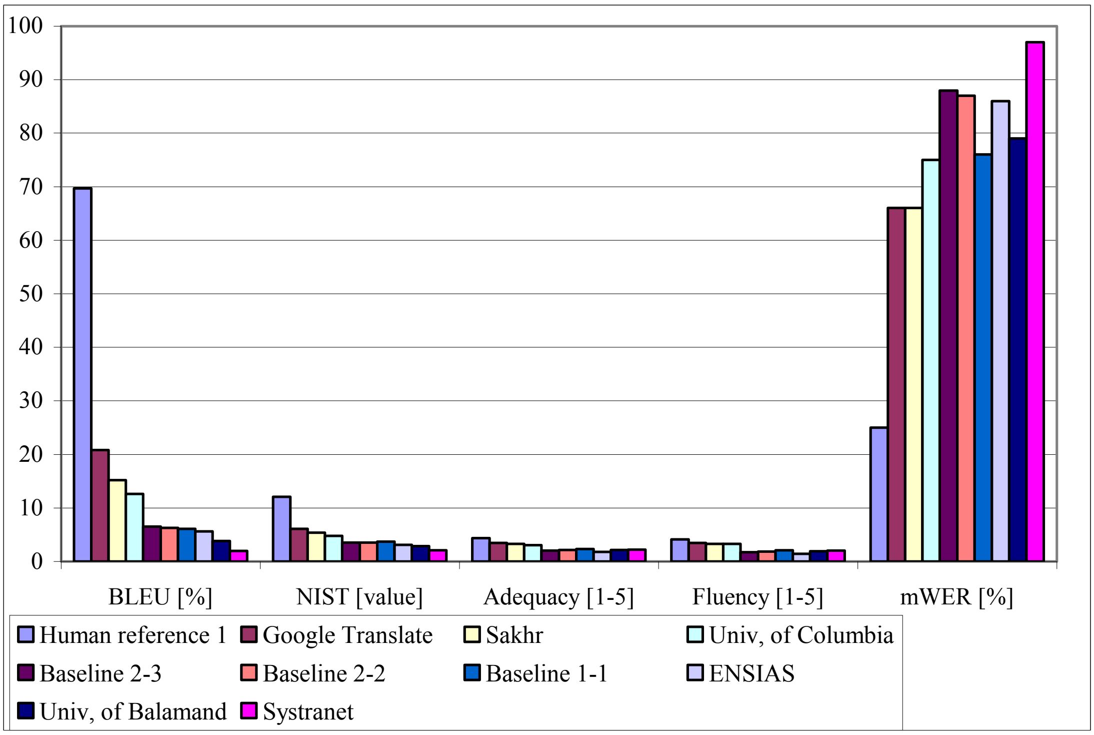

Figure 3: Summary of the automatic and human evaluation results.

Pearson correlation coefficients are presented in Table 22.

	Adequacy	**Fluency**
BLEU	0.87	0.78
NIST	0.91	0.84
mWER	-0.90	-0.83

Table 22: Pearson correlation coefficient on scores between automatic and human metrics.

Meta-evaluation results confirm the automatic metrics work well, but not perfectly. Pearson correlation coefficients are either around 0.80 for fluency, or around 0.90 for adequacy. Here, automatic metrics correlate better with adequacy than fluency, contrary to previous evaluation campaigns. One of our hypotheses about that difference is related to the difficulty to translate complex sentences.

8 Lessons learnt

The general results of this MEDAR evaluation campaign remain stable compared to the dry-run. Although the test data are different, the results of the two online systems allow us to draw this conclusion since their scores did not evolve a lot. However, using training data on the MEDAR baseline system improved the scores, at least by one point of BLEU. 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 and particularly find better parallel training data that fit the vocabulary of the test corpus. This can be accentuated by importing data from several domains and then bringing a large range of lexica. Dealing with out-of-vocabulary can be a complex task, but solutions exists, such as in (Habash, 2008).

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. The post-processing generation is also essential and requires more work for sentence reconstruction (e.g. gender or number)

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 easily adapted to an agglutinative language like Arabic).

4. Improve Moses for Arabic in the same way the University of Balamand did for its own system, such as reordering words for alignment, syntactic analysis for preprocessing, segmentation and morphological decomposition, word alignment, etc.

9 Further work

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. We also want to emphasize that, despite all MT R&D efforts most of the work done on Arabic is on Arabic as a source language.

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

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

10 References

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L. and, Roossin, P.S. (1990), A Statistical Approach to Machine Translation, *Computational Linguistics, Vol. 16,* n. 2, pp. 79-85, 1990.

Brown, P.F., Della Pietra, S.A., Della Pietra, V.J. and, Mercer, M.L. (1993), The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics, Vol. 19, n.* 2, pp. 263-311, 1993.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 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).* Association for Computational Linguistics, Morristown, NJ, USA, 17-53.

Dorr, B. (1993), *Machine Translation*, MIT Press, Cambridge, 1993.

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.

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

Ghaoui, A., Yvon, F., Mokbel, C. and, Chollet, G. (2005) On the Use of Morphological Constraints in N-gram Statistical Language Model, In *Proceedings Interspeech, 9th European Conference on Speech Communication and technology*, pp. 1281-1284, 2005.

Habash, N. (2008) Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*: Short Papers, June 16-17, 2008, Columbus, Ohio

Hamon Olivier, Mostefa Djamel, Arranz Victoria (2008). Diagnosing Human Judgments in MT Evaluation : an Example based on the Spanish Language. In *Proceedings of MATMT, February, 2008*, San Sebastian, Spain, pp 19-26.

Hutchins, W.J. and, Domers, H.L. (1992), *An Introduction to Machine Translation,* Academic Press, Cambridge, 1992.

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, 2007.

Och, F.J. and, Ney, H. (2000), *A Comparison of Alignment Models for Statistical Machine Translation, COLING'00*, pp. 1086-1090, Germany, 2000.

Och, F.J. (2003), Minimum Error Rate Training for Statistical Machine Translation, In *Proceeding of the ACL 2003*, pp. 160-167, 2003.

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

Sagvall Hein, A., Forsbom, E., Tiedemann, J. , Weijnitz, P. , Almqvist, I., Olsson, L.-J. and, Thaning, S. (2002). Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of the 2nd LREC*, *volume V*, Las Palmas de Gran Canaria, Spain, 2002.

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

Uchida, H. (1996), *UNL: Universal Networking Language - An Electronic Language for Communication, Understanding, and Collaboration*. UNU/IAS/UNL Center. Tokyo, Japan, 1996.

Vogel, S., Ney, H. and, Tillmann C. (1996), *HMM-based Word Alignment in Statistical Translation, COLING'96*, pp.836-841, Denmark, 1996.

Weijnitz, P., Forsbom, E., Gustavii, E., Pettersson, E. and, Tiedmann, J. (2004), MT goes farming: Comparing two machine translation approaches on a new domain, In *Proceedings LREC'04, Vol. VI*, pp. 2043-2046, 2004.

11 Annexes

11.1 Annex A. DTD and example of a monolingual corpus

```
<!ELEMENT fileset (doc*)>
<!ATTLIST fileset fileid CDATA #REQUIRED>
<!ELEMENT doc (s*)>
<!ATTLIST doc id CDATA #REQUIRED>
<!ATTLIST doc lang CDATA #REQUIRED>
<!ELEMENT s (#PCDATA)>
<!ATTLIST s id CDATA #REQUIRED>
<!ENTITY ldquo  "?" >
<!ENTITY rdquo  "?" >
<!ENTITY AElig  "?" >
<!ENTITY Ccedil "?" >
<!ENTITY iacute "?" >
<!ENTITY Eacute "?" >
<!ENTITY aacute "?" >
<!ENTITY eacute "?" >
<!ENTITY ccedil "?" >
<!ENTITY deg "?" >
<!ENTITY ordm "?" >
<!ENTITY laquo "?" >
<!ENTITY raquo "?" >
```

Example :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE fileset SYSTEM "medar_monolingual.dtd">
<fileset fileid="MEDAR">
  <DOC docid="1" lang="ar">
    <s id="1">
     Sentence 1
    </s>
    <s id="2">
     Sentence 2
    </s>
    ...
    <s id="n">
     Sentence n
    </s>
  </DOC>
  ...
</fileset>
```

11.2 Annex B. DTD and example of parallel corpus

```
<!ELEMENT fileset (doc* )>
<!ATTLIST fileset setid CDATA #REQUIRED >
<!ATTLIST fileset srclang CDATA #FIXED "EN">
<!ATTLIST fileset trglang CDATA #FIXED "AR">
<!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 fileset SYSTEM "medar_parallel.dtd">
<fileset fileid="MEDAR" srclang=”en” trglang=”ar”>
  <DOC docid="1" genre=”text”>
    <seg id="1">
     Sentence 1
    </seg>
    <seg id="2">
     Sentence 2
    </seg>
    ...
    <seg id="n">
     Sentence n
    </seg>
  </DOC>
  ...
</fileset>
```

11.3 Annex C. DTD and example of an 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>
```

11.4 Annex D. DTD and example of an 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>
```

11.5 Annex E. MEDAR evaluation – translation examples

Source	how bad is said climate change going to get?
Reference	ما مدى سوءِ ما سوف يحدث مما يُحْكى عن تغير المُناخ؟
Translation	كيف يقال التغير المناخي الذهاب إلى ؟
Fluency	3
Adequacy	1
Comments	Too much translation word/word
Source	This can be done in two actions.
Reference	ويمكن إجراء ذلك في إجراءَيْن.
Translation	هذا يمكن خلال سنتين.
Fluency	4
Adequacy	1
Comments	"two actions" is translated in "two years"
Source	In this way, greenhouses both use radiant energy and also save it via their limiting of convection.
Reference	بهذا الأسلوب، فإن البيوت الزجاجية تستخدم الطاقة المشعة وأيضًا تخزنها من خلال تحجيمها للحمل الحراري.
Translation	وبهذه الطريقة ، كلا من استخدام الطاقة تزداد إشعاعا وأيضا بإنقاذه عبر الحد منها .
Fluency	2
Adequacy	2
Comments	There is a problem of words translation (save=help save=prevent).
Source	Forced migration is the most urgent threat facing poor people in developing countries, they argue, affecting some 155 million men, women and children who have had no choice but to flee their homes and seek refuge elsewhere in their own countries.
Reference	الهجرة الجبرية هي أكبر خطر ملح يواجه الناس الفقراء في البلاد النامية، حسب زعمهم، ويؤثر على حوالي 155 مليون رجل، وامرأة، وطفل الذين لم يكن لهم أي اختيار غير أن يفروا من مساكنهم ويطلبوا اللجوء في أماكن مختلفة داخل بلادهم.
Translation	الهجرة القسرية هو الأكثر خطرا الملحة التي تواجه الشعب الفقيرة في البلدان النامية ، فإنهم يقولون ، تؤثر بعض 155 مليونا من الرجال والنساء والأطفال الذين لم يكن لديه خيار سوى على ترك منازلهم والبحث عن ملجأ في أي مكان آخر في بلدانهم .
Fluency	4
Adequacy	2
Comments	The content is preserved, but there is gender and number unconsistencies.
Source	India, for example, could see a drop of 30 to 40 percent.
Reference	الهند، مثلاً، ربما تشهد انخفاضًا ما بين 30 إلى 40 في المئة.
Translation	الهند مثلاً الممكن الانخفاض 03 الى 40 في المئة.
Fluency	5
Adequacy	2
Comments	"30%" becomes "03%"
Source	This should be done in parallel to taking all necessary steps that can

	preserve the current quantity of permanent water flow in Zarqa River Basin and planning to introduce additional water quantities from non-conventional sources (fully treated wastewater
Reference	ينبغي أن يجري ذلك بالتوازي مع اتخاذ كل الخطوات اللازمة التي تستطيع أن تحافظ على الكمية الحالية للجريان الدائم للمياه في حوض نهر الزرقاء والتخطيط لإدخال كميات مياه إضافية من مصادر غير تقليدية (مياه صرف معالجة بالكامل، حصد مياه الأمطار، المياه الرمادية ...
Translation	هذا ينبغي أن يتم بالتوازي إلى اتخاذ كافة الخطوات الضرورية التي يمكنه المحافظة على كمية المياه تتدفق من الحاضر الدائم في حوض النهر والتخطيط لطرح كميات إضافية من المياه من مصادر غير تقليدية (عوملت بشكل تام ، وأن حصاد المياه ، . . . ألخ . . .) في المناطق ا
Fluency	5
Adequacy	3
Comments	Issue regarding the translation of verbs vs noun (water flow versus flow).
Source	However, let us have a look at different kinds of OTEC-
Reference	وبرغم ذلك، دعونا نلقي نظرة على الأنواع المختلفة من تحويل الطاقة الحرارية للمحيط.
Translation	وعلى الرغم من ذلك ، دعونا ننظر لها في مختلف أنواع
Fluency	4.5
Adequacy	3
Comments	The abreviation “OTEC” is missing
Source	The worst news?
Reference	أسوأ الأنباء؟
Translation	إن أسوأ الأخبار ?
Fluency	4
Adequacy	5
Comments	There is an issue with the word “the” and the question mark.
Source	The process of rehabilitation of the Zarqa River Basin is considered as an urgent national demand, even under the current conditions of deterioration that require high investments in rehabilitation.
Reference	إن عملية إعادة تأهيل حوض نهر الزرقاء تـُعتبَرُ مطلبًا وطنيًّا ملحـًّا، حتى في ظل الظروف الراهنة من التدهور التي تتطلب استثمارات ضخمة في إعادة التأهيل.
Translation	عملية الإصلاح بحوض نهر يعتبر كذلك مطلب وطني عاجل ، حتى في ظل تدهور الأوضاع الحالية لأنه يتطلب استثمارات عالية في إعادة التأهيل .
Fluency	5
Adequacy	4.5
Comments	The proper name “Zarqa river” is missing.