MEDAR

Mediterranean Arabic Language and Speech Technology

Deliverable D5.1.

WP5 work plan

Author: Khalid Choukri, ELDA, Olivier Hamon, ELDA
Contributors: All partners
December 2008

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 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

http://www.medar.info, email: nemlar@hum.ku.dk

CONTENT

1. Executive Summary

This report elaborates on the MEDAR findings related to MT and CLIR/MLIR tools. It describes the objectives of the project which target the selection of a number of tools, their adaptation/customization to Arabic (either as a source or a target language), their evaluation and the development of new Language Resources for their improvement and enhancement. The report elaborates on a detailed work plan for the consortium to produce an SMT baseline and then an enhanced version that would benefit from contributions of all partners. The report also elaborates on the development of an evaluation framework that would help evaluate both the baseline and the improvements achieved. The work plan will distinguish two phases: Phase 1 will allow us to implement the baseline exploiting some of the well know open sources SMT like Moses and Phase 2 will allow us to bring in more of the know-how of the consortium. In both phases language resources will be developed to train and assess the system. The work is distributed over the partners according to the expertise of each as initially foreseen. The project will mostly focus on MT and leave out the CLIR option.

2. Objectives of MEDAR regarding MT & CLIR/MLIR

The MEDAR projects exploits inputs from the surveys conducted at an early stage and reported on the deliverable D3.1 of the project. Such input will help us define the needs in terms of Language Resources, basic tools and components and the corresponding milestones to "customize" identified Machine Translation and other tools for translation and information retrieval. Such tools and components could be both in the open source sphere and within the partners of the project. This deliverable will also identify language resources and tools that are missing (‘gaps’ as identified by the survey), and will list and specify a subset of these for testing hypotheses that have to be produced at a later stage. Evaluation framework will be established and exploited on the basis of partners' expertise, in particular on the basis of ELDA evaluation platforms for MT & CLIR.

The survey conducted within MEDAR has identified a number of players and their expectations and requirements. The respondent to the survey also expressed their views on language resources and tools that are needed (‘gaps'). Unfortunately the survey itself did not collect any information about existing translation and Information retrieval tools. So ELDA and the project partners carried out an identification task and gathered information about tools both for MT and CLIR/MLIR.

The objectives of this task would be:

- To list all identified tools for MT & CLIR/MLIR that process Arabic and that would be made available to the project,
- To short list those that can be adapted to the needs of the project consortium (customization of existing open source as well as proprietary tools)
- Specify and develop a small subset of adequate resources for testing the performance of these systems.

3. MEDAR Requirements

This deliverable aims at defining a work plan on how to select the best tools, to customize them for Arabic (either as a source or a target language) and assess their performances. In order to do so, a first phase will consist of an "exhaustive" identification of existing tools. Such work will be based on the consortium knowledge rather than on the survey. Major sources of input could be projects like CLEF (and the current Treble-CLEF), TREC, Euromatrix and its MT Marathon, etc.

The work will also help the consortium investigate the best solution for the combination(s) of existing modules to offer guidelines for the development of technologies for Arabic, based on components that exist as open source code or as background knowledge of partners. The consortium would also like to work on the enhancement of some of the tools through the use of specific resources and tools that would be identified (e.g. the exploitation of corpus alignment work for which partners have valuable expertises).

The other task to be carried out within this work package is the definition of a stable framework for MT evaluation and improvements. Several methodologies exist for the evaluation and benchmarking of MT. As indicated above, the project will build on the state-of-the-art and previous experiences of the project partners to define a baseline MT system (and the corresponding components) exploiting partners' tools willing to participate and (when available) other participants and open source software. These tools also have to be used to validate the evaluation framework as long as this is not the one at hand but derived from it. Metrics will be based on existing automatic evaluation methodologies. The consortium may also benefits from the expertise acquired by ELDA in other projects (TC-STAR, CESTA) to carry Human evaluation as well as user centered evaluations.

The MT tools will be assessed at the beginning and after the inclusion of resources and tools developed by the project, and comparison will be made.

In order to carry out extensive evaluations of MT, no specific evaluation work is planned for CLIR. Close connections are foreseen with CLEF to discuss potential evaluation in the CLEF framework.

The tasks of this WP are:

- Decide on which MT and CLIR/MLIR tools to consider, and which type of LR or tools may improve performance
- Set up an evaluation framework and evaluate the baseline system (Set up evaluation methodology, test sets etc.)
- Produce the LRs needed to improve the baseline
- Update the tool
- Perform evaluation after the use of additional material

4. Some other general issues to consider

It is important for the project and the HLT community interested in Arabic to develop/customize existing tools so as to obtain a well recognized state of the art baselines. MEDAR is handling this need for MT and for CLIR/MLIR.

In order to do so a number of issues have to be considered:

- The standardization of input and output (though we will recommend to use Unicode), it is important to ensure that the tools selected can handle that.
- The language variety: It is important to focus on the Modern Standard Arabic but one should not leave out colloquial Arabic(s) if possibilities are offered to do so.
- The availability of the baselines after the customization

5. Conclusion of the MEDAR Survey results

The survey and the partners collected information about existing HLT applications and those who were indicated are listed in the project deliverable D3.1. These applications/technologies were not prioritized though MT, CLIR/MLIR and ASR are mentioned many times which consolidate the project original plans to focus on MT and CLIR/MLIR and if manpower permits on Spoken Document Retrieval that combines speech processing with CLIR/MLIR.

MT findings

The first result of this information collection is given in the following matrix that lists all identified MT systems which includes Arabic either as a source or a target language. We did not manage to find information about the components of the systems as many developers just reported that this was internal technology (or an adaptation/deployment of a known one e.g. Systran by Google). We have distinguished the following general features:

Product	The name given to the "system"
Type	Profile of the owner (commercial versus academic), this does not help to draw any conclusion regarding availability of the software.
Owner of the "product"	The company or the university that owns the system.
Nature (TM, SMT, Hybrid, RB, unknown)	The technology behind the system (Translation Memory, Statistical Machine Translation, Rule-based, a combination of the two hybrid, or the information is not know)
Basis of the system	Whether the technology is developed internally or it uses well identified components for other parties.
Availability	Means availability for end-users (not necessarily availability of the system as open-sources or modules for developers)

Product	Type	Owner of the "product"	Nature (SMT, Hybrid, RB, unknown)	Basis of the system	Availability	Ar->En	Ar->Fr	Ar->Es	En->Ar	Fr->Ar	Es->Ar
Al Misbar	Commercial	ATA Software Technology Ltd.	Unknown	Internal	Free (Web)				x		
Al Mutarjim Al Arabey	Commercial	ATA Software Technology Ltd.	Unknown	Internal	£200	x			x		
Al-Wafi	Commercial	ATA Software Technology Ltd.	Unknown	Internal	£50	x			x		
Angusman's Translator Plugin Pro	Commercial	Taragana	Unknown	Unknown	$30				x		
An-Nakel El-Arabi	Commercial	Cimos	TM	Internal	>$500	x	x		x	x	
Applied Language Solutions	Commercial	Applied Language Solution	Unknown	Unknown	Free (Web)	x			x		
ArabTrans	Commercial	ArabNet Technology	Unknown	Unknown	Unknown				x		
BBN Technologies	Commercial	BBN Technologies	SMT	Internal	Unknown	x					
Golden al-Wafi	Commercial	ATA Software Technology Ltd.	Unknown	Internal	£75	x			x		
Google Translate	Commercial	Google	SMT	Internal	Free (Web)	x			x		
IBM Statistical Machine Translation System:	Commercial	IBM	SMT	Internal	Service Offering from IBM	x			x		
Interpret	Commercial	Interpret	Unknown	Unknown	Free (Web)	x					
Language Weaver SMTS	Commercial	Language Weaver Inc.	SMT	Internal	Unknown	x	x	x	x	x	x
LEC series (Passport Premium, Translate DotNet, etc.)	Commercial	Language Engineering Co.	Unknown	Internal	>$750	x	x	x	x	x	x
Maximum Edge	Commercial	MaximumEdge.com	Unknown	Unknown	Free (Web)	x			x		
MITRE Corporation	Commercial	MITRE	Unknown	Unknown	Unknown	x					
MTM Linguasoft Online Translation	Commercial	MTM Linguasoft	Unknown	Unknown	Free (Web)	x			x		
MutarjimNet	Commercial	ATA Software Technology Ltd.	Unknown	Unknown	Unknown	x			x		
Sakhr Enterprise Translation	Commercial	Sakhr Software Co.	TM	Internal	Unknown	x			x		
Systran series	Commercial	Systran Co.	RB/Hybrid	Internal	>$100	x			x		
Tarjim	Commercial	Sakhr Software Co.	Unknown	Internal	Free (Web)	x			x		

Transclick	Commercial	Transclick Inc.	Unknown	LEC	$5/month	x			x		
Translate-Net	Commercial	Cimos	Unknown	Internal	$990	x	x		x	x	
Translution series(Business, for enterprise, etc.)	Commercial	Translution	Unknown	Internal	Unknown	x			x		
TranSphere	Commercial	AppTek Inc	TM	Internal	Unknown	x			x		
WebTrans	Commercial	AppTek Inc	Unknown	Internal	Unknown	x			x		
Windows Live Translator	Commercial	Microsoft Corporation	SMT	Internal	Free (Web)	x			x		
Carnegie Mellon	Academic	Carnegie Mellon	SMT	Internal	Unknown	x					
Fitchburg State College	Academic	Fitchburg State College	Unknown	Unknown	Unknown	x					
Johns Hopkins University	Academic	Johns Hopkins University	SMT	Internal	Unknown	x					
Queen Mary University of London	Academic	Queen Mary University of London	SMT	Internal	Unknown	x					
RWTH University of Aachen	Academic	RWTH University of Aachen	SMT	Internal	Unknown	x					
Technical University of Catalonia (UPC)	Academic	Technical University of Catalonia (UPC)	SMT	Internal	Unknown	x					
U.S. Army Research Laboratory	Academic	U.S. Army Research Laboratory	Unknown	Internal	Unknown	x					
Université du Maine	Academic	Université du Maine	SMT	Internal	Unknown	x					
University of Cambridge	Academic	University of Cambridge	SMT	Internal	Unknown	x					
University of Edinburgh	Academic	University of Edinburgh	SMT (not Moses)	Internal	Unknown	x					
University of Maryland	Academic	University of Maryland	SMT	Internal	Unknown	x					
University of Southern California, Information Science Institute	Academic	University of Southern California, Information Science Institute	SMT	Internal	Unknown	x					

A number of tools are also considered important by the community; these were also identified and are listed herein:

Type	Name
Morphological Analyzers	ArabMorpho
	Xerox Arabic Morphological Analyzer
	Raramorph
	Buckwalter Arabic Morphological Analyzer
	Sebawai
	Morphological Analyzer (CRL, New Mexico State University)
Stemmer	Al-Stem
	Light10
	Larkey
POS Tagger	ArabTagger
	MorphTagger
	Stanford Log-linear Part-Of-Speech Tagger
	Brill's POS tagger for Arabic
Parser	Stanford Arabic Parser
Statistical Machine Translation Toolkit	Egypt
Syntactic Analyzer	Syntactic Analyzer (Cimos)

Finally, ELDA identified a number of LRs that could be used to train the selected tools or to better tune them to Arabic and the given domains:

Type	Name
Dictionaries	Al-Misbar
	Al-Wafi Quick Dictionary
	ATA-NTS
	Ajeeb
	Ectaco
	Babylon-Pro
	FreeDict
	LingvoSoft
	Pan Images
	Partner
	PocketTran
	TranslationBooth
	WordPoint
	Xpro7

	ArabDictions Monolingual lexicon Arabic full-form lexicon
Bi/Multilingual Lexica/Dictionary	Sakhr Multilingual Dictionary
	DixAF (Ar-Fr, ELRA)
	Arabic-Spanish Verbs based lexicon
	Modern Standard Arabic-Dutch dictionary
	CRL Arabic-English Dictionary
	UB Diccionari Arab <> Anglès <> Castellà Modern Standard Arabic-Dutch dictionary
Parallel Corpora	UN Bidirectional Multilingual (En, Fr, Ar, Ru, Zh)
	UNESCO
	Hebrew-Arabic-English corpus (Agava Institute)
	EGYPT Gizza Toolkit Quran Parallel Corpus (Ar-En)
	CLARA (Corpus Linguae Arabicae) (Ar-Cz)
	Bilingual aligned corpora (Ar-It, ILC)
	Umaah Arabic English Parallel News Text
	Arabic-English Parallel Translation (LDC)
	10k words AFP Arabic Newswire corpus translated into English (LDC)
	Euradic (Ar-Fr)
	E-A Parallel Corpus (University of Kuwait)
Bilingual Corpora	Multiple Translation Arabic (LDC)
	TDT4 Multilanguage Corpus (LDC)
	STRAND English-Arabic Parallel Web Pages (Tool and a corpus)
Monolingual Arabic corpora	An-Nahar Newspaper (ELRA)
	Arabic Data Set (ELRA)
	Le Monde Diplomat ique (ELRA)
	NEMLAR Written Corpus (ELRA)
	DIINAR.1 (ELRA)
	AFP Arabic corpus (ELRA)
	Al-Hayat Arabic corpus
	Nijmegen Corpus
	ArabiCorpus 11,000 arabic Wikipedia Articles (Benajiba)
Evaluation corpora	Arcade II Evaluation Package (Le Monde Diplomatique corpus containing aligned sentences for Arabic-to-French – 316 000 words)
	CESTA Evaluation Package (The two corpora from Le Monde Diplomatique and from the UNICEF, WHO and FHI websites – 60 000 words translated from 1 to 4 times)

The next task is to elaborate a detailed description framework (metadata) to help us decompose each of the identified tools and also assess how they comply with our selection criteria.

CLIR/MLIR findings

Regarding the CLIR and/or MLIR, ELDA has identified the following tools and resources. A deep analysis of CLEF & TREC campaigns will be conducted to obtain more data.

Tool	**Product**
Text Search Engine	Swift
	Google
	Yahoo
	4Arabs
	Ayna
	Arabo
	Yamli
	MSN
	Exalead
	Araby
Question Answering	AQAS

Resources (many are identical to the MT findings):

Type	**Name**
Monolingual Corpo	Agence France Presse (LDC, ELRA also comparable/parallel corpus)
	Al-Hayat
	Arabic corpus (Leeds)
	Leuven Corpus
	Nijmegen Corpus
	DINAR corpus
	General Scientific Arabic Corpus
	Classical Arabic Corpus
	SOTETEL
	Corpus of Contemporary Arabic
Treebank	Penn Arabic Treebank
	English-Arabic Penn Treebank
	Prague Arabic Treebank
Topics, Queries, Questions	200 Questions of different types (Benajiba)
	CLEF Questions in Arabic
	Arabic Translations of TREC 2001 IR topics

Like the plans for MT, a next task is to elaborate a detailed description framework (metadata) for CLIR/MLIR to help us decompose each of the identified tools and also assess how they comply with our selection criteria.

6. MT principles and selection of the short list

After our review of the state of the art of MT technologies, we have opted for a number of available tools, either within the community (e.g. Open source tools) or within the partners. Among the requirements for SMT tools that we have considered, we have:

- The main bottleneck is to get the parallel corpus for e.g. Moses that is language neutral and open source.
- Ensure that a large corpus is available and ensure that it is not tied to a particular domain (e.g. we can get a web corpus from UNESCO, UN, but it will be domain specific).
- The amount of Arabic data has to be much bigger than for other languages, because of the morphology/word formation unless to consider Arabic stems to overcome the sparseness (which requires lot of morphological analysis and a POS-tagging).
- We should also consider for the evaluation purpose a Rover (or system combination, e.g. rule based and statistical (may be we could do that with the systems of Sakhr and IBM (as black boxes).
- We can use alignment tools produced by the partners to improve the quality of the parallel corpus and thus use less data for comparable performance.

With these considerations in mind, the consortium has agreed to adopt Moses as the main SMT kit and its baseline.

7. Selection of an SMT solution and MEDAR related tasks

The MEDAR proposition for an SMT

We focus in this paragraph on the description of MT modules, in particular on the Statistical MT ones (SMT). The consortium agreed to use a state of the art package called Moses.

The rationales behind using Moses in this context are:

- Moses is an Open Source package
- Proven to be of identical quality to proprietary state of the art MT systems.
- Has been used for several other language pairs.
- It may include easily other linguistic information (factored translation model).
- It has advantages for speech-to-speech translation as it may take as input the results of automatic speech recognition (not only the best solution of a recognizer but we may provide a network of solutions (confusion network decoding)).

The details about SMT and Moses can be found at:
http://www.statmt.org/Moses/?n=MOSES.Background

The partners will also consider the use of other toolkits for comparison such as: the GenPar (Toolkit for Research on Generalized Parsing, (including Machine Translation by Parsing) which also provides an architecture, a design, and an implementation of an integrated system for statistical machine translation by parsing (more details at:
http://nlp.cs.nyu.edu/GenPar/GenPar.html)

The Moses Components that will have to be customized for our project are:

- Language resources and data preprocessing (e.g. the language model should be trained on a corpus that is suitable to the domain, preferably a parallel corpus)

- Language Modeling toolkit , here we can choose one of the following kits:
    - the SRI language modeling toolkit, which is freely available.
    - the IRST language modeling toolkit, which is freely available and open source.
    - the RandLM language modeling toolkit, which is freely available and open source.

  And very likely the first one (SRILM for language modeling),

- GIZA++ for word alignments

- Tuning the translation models (minimum error rate training) but also exploit a number of features such as:
    - Reorder phrases and lexicons
    - First pass of translation using Moses generating n-best (e.g. n=1000)
    - Second pass reordering the n-best solutions with a more precise language model

The results can be then evaluated using:

- An automatic metric like BLEU

- Language resources that consist of translated texts (called "reference translated texts" with at least 4 different translations)

In the work plan we assumed that there will be two phases:
1) Building a baseline

2) Improving the baseline both with new Language Resources (new parallel corpus) and new tools (e.g. word alignment tools using partners' multilevel aligners).

The different phases and tasks

The work plan defined within the technical annex of the project schedule the work as:

	Duration in months																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
WP5																														

The associated resources are:

WP No	Work package title	Type of activity	Lead partner	Person-months	Start month	End month
5	Tools for translation and information retrieval for Arabic, resources and evaluation	Support Activity	ELDA	57	8	29

The original deliverables planned within this activity were:

Del. no.	**Deliverable name**	**WP no.**	**Lead bene-ficiary**	***Estimated indicative person-months***	**Nature**	**Dissemi-nation level**	**Delivery date[1] (proj. month)**
5.1	WP5 work plan	5	2	1	Report	Public	10
5.2	Language resources and tools for Arabic MT and IR	5	2	36	Other	Public	24
5.3	Evaluation of tools for Arabic, recommendations	5	2	20	Report	Public	29

1 Month in which the deliverables will be available. Month 1 marking the start date of the project, and all delivery dates being relative to this start date.

The manpower associated to this activity is given within the following table. This covers all activities including the preparation of this workplan:

Part. no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
Part. short	UCPH	ELDA	UOB	AU	UU	ILSIP	RDI	BIT	ENSIAS	CEA	CNRS	OU	Lyon2	IBM	Sakhr
PM	4	5	6	6		4	6	2	6	6				6	6

This work package will be split into two phases. The first one (Phase 1) will ensure that a baseline is developed and evaluated, while the second one (Phase 2) will ensure that contributions from partners will ensure a substantial enhancement of the baseline performances. Such contributions will cover booth language resources and the various tools used to optimize the translation module (e.g. multi-level alignments).

The outcome of the work will constitute deliverables D5.1, D5.2 and D5.3. The deliverable 5.1 is this document while D5.2 and D5.3 will be delivered in two versions: one corresponding to Phase 1 (D5.2a and D5.3) and the second to phase 2 (D5.2b and D5.3b).

Tasks of phase 1

The objective of this task is to build a baseline to be shared by all partners and widely disseminated. The tasks to carry out are:

1. Build a parallel corpus Arabic ⇔ English (and may be other languages if feasible at low cost), this can be achieved by identifying data within multilingual content producers (UN, Unesco, etc.)
2. Align the corpus using the Giza++ aligner
3. Collect a huge monolingual corpus for Arabic and English to train the language models
4. Install and run Moses (train its decoder), exploiting its various features
5. Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus)
6. Evaluate the whole system using BLEU (and other automatic metrics) and ensure that it is compared to systems brought by the partners (in particular SMT systems).

Tasks of phase 2

Phase 2 will depend heavily on the performance of phase 1 and will be planned in details afterwards. In particular two options will be considered: a) increase the size of the language resources to train the tools versus b) change the domain/genre of the data to see how robust the system is to new domain/genre. The objective is to identify which components can/should be adapted, customized and enhanced for Arabic and which resources and tools are needed to enhance the baseline. The main tasks will be:

1. Build a parallel corpus Arabic ⇔ English (and may be other languages if feasible at low cost), either to enrich what has been used in Phase 1 or to cover a new domain (health, economics, etc.)
2. Align the corpus using some of the partners aligners in addition to the Giza++ aligner
3. Collect a new and huge monolingual corpus for Arabic and English to train the language models (only) if we feel that the domain is so different that it requires a new language model
4. Consider the possibility to use new features within Moses like exploitation of morphological analyzed corpus (alignment of Pos)
5. Run Moses with the new datasets
6. Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus if these were not used during phase 1)
7. Evaluate the whole system using BLEU (and other automatic metrics)

8. MT Evaluation framework for MEDAR

As indicated in the project proposal, MEDAR will focus on MT for the evaluation of performances. To do so, the project will specify and develop a small subset of adequate resources for testing the performance of these systems. This will be carried out in a stable framework for MT evaluation that will be adapt by the project on the basis of evaluation campaigns carried out by ELDA.

In addition to the project baseline MT system (and the corresponding components), partners (and beyond that the MEDAR/NEMLAR network of players) will be encouraged to bring in their own tools for comparisons.

The MT tools will be assessed at the beginning and after the inclusion of resources and tools developed by the project, and comparison will be made to identify improvements as indicated both in the tasks of Phase 1 and Phase 2.

Once the project has decided on which MT tools, and which type of LR or tools may improve performance, an evaluation framework (as a web service) will be set up and made widely available. Depending on the number of tools/systems we may re-use resources from former projects like Evalda-CESTA.

It is crucial to evaluate both the baseline system and the enhanced version.

Evaluation methodology (from CESTA, TC-STAR, NIST-MT)

The MEDAR MT evaluation will be organized considering Arabic as either a source or a target language. MEDAR does not intend to carry R&D activities on MT evaluation but rather take inspiration from previous projects conducted by the partners such as Evalda-Cesta[2]or TC-STAR[3].

The first phase of this action is to identify existing MT tools and the ones which can be adapted and/or customized to Arabic. In addition to the tools that would help us design MEDAR MT system(s), we will encourage all MEDAR partners who own a MT system with Arabic as one of the languages to join. This invitation will be extended to players outside the project (many of them participated to CESTA that handled French to Arabic/English pairs). Once we have a clear picture of who wants to participate, we will have to define an evaluation protocol which includes human and automated quality metrics, and to assess its reliability on the EN/FR and EN/AR language pairs.

Two runs will be organized. The first one aims at evaluating output quality, from absolute and comparative points of view, on a general-domain reference corpus and the baseline system adapted within MEDAR. The concrete details of the evaluation protocol will be reviewed and if necessary revised/improved after the first run. The second run aims at measuring the capacity of systems to adapt to a new domain in a very limited amount of time. The project will develop a set of Language resources and participants will receive such “adaptation” data for a specific domain as will be explicitly defined within the tasks of Phase 2.

The idea is to compare the performance before and after using the adaptation data to improve the systems’ performances and enrich its features.

Systems can be anonymized if necessary using e.g. CESTA conventions for this purpose.

Evaluation Metrics that could be used

The MEDAR project is considering the use of human judgments of fluency and adequacy as the reference for translation quality levels in addition to the usual automatic measures. An evaluation platform that automates such process has been developed and deployed by ELDA in previous projects and can be adapted to MEDAR needs. The “quality” of automated evaluation metrics can therefore be assessed with respect to human scores, checking whether automatic scores reproduced the human ones or at least the rankings derived from them. MEDAR can profit from the CESTA protocol that resembled the one used by the National Institute of Standards and Technology (NIST 2003).

Let us elaborate quickly on what was used within CESTA and which could be reused within MEDAR. The two CESTA runs included human evaluations of the quality of the output of MT systems, which represented the most costly measure of the campaign. Two well-known parameters of translation quality were assessed, namely fluency (or intelligibility) and adequacy

[2] CESTA stands for *Campagne d’Évaluation des Systèmes de Traduction Automatique*, and was supported by the Technolangue program of the French Government (details www.elda.org/cesta)

[3] Reference to TC-STAR (www.tc-star.org)

(or fidelity), following the DARPA 1994 campaign (White et al. 1994). These parameters, together with more detailed alternatives, are given major importance in the FEMTI guidelines for MT evaluation (Hovy et al. 2002).

The human judgments were obtained using our evaluation platform which offers a Web-based interface (Hamon et al. 2006) which displays translated segments (generally sentences) to the users in a random order, so that quality judgments are kept as independent as possible between adjacent segments. Each segment was evaluated by two judges.

In addition to human judgments, CESTA employed several automated metrics. Three of them, referred to as BLEU, NIST and WNM, are based on a comparison between the candidate translation and one or more reference translations. These are the ones we suggest to reuse within MEDAR.

As a brief reminder, the BLEU metric (Papineni et al. 2001) and its NIST variant (Doddington 2002) make use of n-gram based comparisons ($n = 1.5$) between the candidate translation and, typically, up to four reference translations. The more n-grams the candidate segment has in common with the reference segments, the higher the score, and a penalty is introduced for much shorter candidates. The number of unigram matches has been shown to emulate fidelity scores, while higher-level n-gram matches are closer to fluency scores.

The WNM metric (Weighted N-gram Model) (Babych et al. 2004, 2005) refines the n-gram based comparison by weighting the words according to their importance, which is computed using a variant of the *tf.idf* score used in information retrieval. WNM also defines recall, precision, and f-measure; its authors show that recall emulates human judgments of adequacy, while the f-measure best corresponds to human fluency scores.

Although automated metrics have been shown to have some limits (e.g. a score increase does not always reflect an increase in translation quality), their relatively low application cost makes them widely used.

Organization of the Two MEDAR Runs

The CESTA evaluation campaign was scheduled as two runs, which were conducted at an interval of some 12 months (data for the second run was produced in parallel to the execution of the first run). These two test runs were preceded by a dry run designed to test the integrity of the data distribution and processing systems hosted by the CESTA organizers. We plan to re-use a similar approach within MEDAR except that the interval between the two runs will be reduced to 3-4 months.

8.1.1. First Run

The goal of the first run is to provide an initial measure of the quality of the MT baseline and any participating systems, using the full set of metrics selected for CESTA and texts taken from the general domain. The general domain remains to be defined as well as the quantity of the data.

8.1.2. Second Run

In addition to the comparative evaluation of system performance, one of the goals of the second run was to assess the capacity of the systems to adapt or to be customized to the subject domains of the source texts. For this reason, an improved evaluation protocol as designed by CESTA, exploited two sets of scores, one using a 'generic' or 'default' version of the systems being evaluated and the other using versions customized for the subject area chosen.

Depending on the efforts that will be estimated within the work plan, the MEDAR consortium may decide to go for these two runs or keep the simple one.

9. Potential contribution of Partners to the development of the MEDAR baseline

As a basis we imagine that each partner would contribute with its expertise and background, in particular partners who have allocated manpower for this purpose. This is a first attempt and detailed work will be specified at a later stage (once the short list and the background tools are known to all).

UCPH suggestions and Plans

UCPH has a valuable expertise in Evaluation and can contribute in helping ELDA to carry out the evaluations.

ELDA suggestions and Plans

ELDA has extensive expertise in identifying existing LRs and developing new ones as well as carrying out Evaluation of Human Language Technologies. ELDA can contribute with the specification and production of LRs to adapt/customize a baseline MT system to Arabic/French and/or Arabic/English. ELDA can also take in charge the evaluation of MT on the basis of the protocol and methodology described above.

UOB suggestions and Plans

UOB has extensive expertise in speech processing and has several components such as Multilingual speech recognition applied on Broadcast news (bilingual), Speaker recognition, Handwritten recognition, Morphological based statistical language modelling, Audio Visual speaker recognition, Indexing. The tools developed within the team are: BECARS: a GMM freeware: www.tsi.enst.fr/becars, HCM: HMM toolkit, UOB-CART: CART algorithm, UOB-MLP: MLP. The Freeware tools we adopted and sometimes extended are: HTK, SPHINX, SRILM.

UOB can contribute to the Spoken Document retrieval within MEDAR and/or the design and production of LRs for evaluation as it did within NEMLAR.

AU suggestions and Plans

AU can help in three potentials areas:

- The exploitation of the UNL approach and how it is applied to Arabic language (based on the work carried out by Dr. Daoud Maher, PhD from Geta-Grenoble, France).
- A new spell checker is developed and released which can be used as a component of SMT tools. It is based on a hybrid approach which is utilizing morphological knowledge in form of consistent root-pattern relationships, and some morpho-syntactical knowledge based on affixation and morpho-graphemic rules to specify the word recognition and non-word correction process.
- Arabic Speech Recognition system using Sphinx V engine, a newly developed system. This can be done if MEDAR addresses the Spoken Document retrieval issue.

ILSP suggestions and Plans

ILSP suggested going for "**an all-encompassing target**", namely multilingual retrieval from spoken documents. Such a target would give us the opportunity to:

- integrate three important pillars : machine translation, information retrieval and speech processing/recognition
- integrate all the expertise available in the Arabic partners' sites

As for expertise relevant to the above, ILSP can contribute with:

- multi-level parallel text processing, from sentence to word/term alignment,
- translation memory and statistical machine translation, using ILSP own tools
- its experience with open source tools

RDI suggestions and Plans

RDI has developed a large number of tools for written/spoken Arabic. It may contribute with:

- Arabic morphological analyzer, Part-of-Speech tagger, and Lexical Semantic tagger (including Statistical based ones).
- HMM-based Arabic font written OCR which may help collect more corpora.

These tools may be used within the project but will probably not be put in the open source domain nor given to the partners after the end of the project (a free license excluded).

BIT suggestions and Plans

BIT has some expertise and an important know how in Arabic linguistics and will contribute to the production of the parallel corpus. They will also carry out tasks to ensure that the alignment is validated.

ENSIAS suggestions and Plans

ENSIAS has a strong background in speech processing and recently conducted some work in partnership with ELDA on adapting existing open source speech recognition systems to Arabic. The work focused on the CMU Sphinx and HTK tools. Such work could be continued if the project addresses the Spoken Document retrieval.

ENSIAS can also help with the design and production of LRs for evaluation as it did within NEMLAR.

CEA suggestions and Plans

CEA can contribute to the improvement of **a** baseline MT system and a baseline MLIR system using:

- their bilingual lexicons with word alignment techniques,
- their linguistic components (tokenization, morphological analysis, pos tagging, syntactic analysis).

Among the existing building blocks at CEA we can use a word aligner which can be used to build bilingual lexicons from parallel corpora.

IBM suggestions and Plans

IBM may develop a baseline SMT system on the basis of state of the art open source and introduce enhancements over baseline using resource from partners (e.g., alignments, lexical dictionaries, etc). IBM can provide segmentation, POS tagging and Named Entity tagging to the corpus that will be used during the course of building SMT starting from an open source system.

SAKHR suggestions and Plans

SAKHR will use its own MT system and compare it to the ones being developed within the project. Sakhr may benefit from the components of the baseline system to improve its own (e.g. the corpus aligner).

10. The Work plan to develop the MEDAR MT for Arabic

Work plan of Phase 1

10.1.1. List of the Phase 1 tasks

Task Id	Tasks
T1	Build a parallel corpus Arabic ⇔ English (and maybe other languages) a. Specification b. Identification of existing ones c. Formatting, cleaning d. Clearing IPR issues
T2	Align the corpus using the Giza++ aligner: a. Install the Giza++ at all involved sites b. Adapt and Run Giza++ for Arabic/English
T3	Collect a huge monolingual corpus for Arabic and English to train the language models
T4	Install and run Moses (train its decoder), exploiting its various features
T5	Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus)
T6	Evaluate the whole system using BLEU (and other automatic metrics) and Human Evaluations

10.1.2.Involvement of different partners is tasks of Phase 1

Part. no	**Part. no**	1	2	3	4	6	7	8	9	10	14	16	**Global manpower**
Part. short	**Part. short**	UCPH	ELDA	UOB	AU	ILSP	RDI	BIT	ENSIAS	CEA	IBM	Sakhr	**All partners**
T1	Build a parallel corpus Arabic ⇔ English (and may be other languages) a. Specification b. Identification of existing ones c. Formatting, cleaning d. Clearing IPR issues	0.5	1	0.5	3	0	3.5	1	0	0	0	0	**9.5**
T2	Align the corpus using the Giza++ aligner: c. Install the Giza++ at all involved sites d. Adapt and Run Giza++ for Arabic/English	0	0.5	1	0	1	1		0	1.5	1	1	**7**
T3	Collect a huge monolingual corpus for Arabic and English to train the language models		0.5	0	0		0	1	0	0	0	0	**1.5**
T4	Install and run Moses (train its decoder), exploiting its various features		0.5	3	0.5	1.5	0		3.5	0	1	1	**11**
T5	Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus)	1	0.5	0	0.5		0		0	0	0	0	**2**
T6	Evaluate the whole system using BLEU (and other automatic metrics) and Human Evaluations	1.5	0.5	0.5		0	0		0	0	0	0	**2.5**
PM	**Manpower/task**	**3**	**3.5**	**5**	**4**	**2.5**	**4.5**	**2**	**3.5**	**1.5**	**2**	**2**	**33.5**

10.1.3.Planning of Phase 1

Phase 1 is planned to last 12 months (January 2009-December 2009).

The planning is given herein:

	Duration in months																													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Phase 1												1	2	3	4	5	6	7	8	9	10	11	12							
Phase 1-T1																														
Phase 1-T2																														
Phase 1-T3																														
Phase 1 – T4																														
Phase 1- T5																														
Phase 1- T6																														

10.1.4.Deliverables of Phase 1

The deliverables planned within this work package are:

- *D5.2 will comprise list of MT and IR tools and customized versions, appropriate resources (parallel corpora Arabic-English, Arabic-French or other language pairs), Test sets for the evaluation, and delivered at T24.*

- *D5.3 Evaluation methodology and results. Recommendations for the future. Delivered at T29*

So phase 1 will produce deliverable D5.2 (that will be also amended at the end of Phase 2) and will contribute to the deliverable D5.3. These will consist of:

(i) A fully operational SMT for Arabic ⬄ English.
(ii) A Language resource kit to train and evaluate the system.

Work plan of Phase 2

10.1.5. List of the Phase 2 tasks

Task Id	Tasks
T1	Build a parallel corpus Arabic ⇔ English either to enrich what has been used in Phase 1 or to cover a new domain (health, economics, etc.) a) Specify the objective (extend the data or new domain) b) Identify adequate resources c) Collect and format them d) Clear IPR
T2	Align the corpus using some of the partners aligners in addition to the Giza++ aligner
T3	Collect a new and huge monolingual corpus for Arabic and English to train the language models
T4	Consider the possibility to use new features within Moses like exploitation of morphological analyzed corpus (alignment of Pos): a) Define new features available to the project b) Produce the right resources with such features (e.g. PoS)
T5	Run Moses **and the partners SMTs** with the new datasets
T6	Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus if these were not used during phase 1) a) Check validity of CESTA Corpus b) If not adequate: specify, collect and translate a new test set
T7	Evaluate the whole system using BLEU (and other automatic metrics) and Human Evaluations

10.1.6.Involvement of different partners is tasks of Phase 2

Part. no	Part. no	1	2	3	4	6	7	8	9	10	14	16	Global manpower
Part. short	**Part. short**	UCPH	ELDA	UOB	AU	ILSP	RDI	BIT	ENSI AS	CEA	IBM	Sakhr	**All part**
T1	Build a parallel corpus Arabic ⇔ English either to enrich what has been used in Phase 1 or to cover a new domain (health, economics, etc.)		0.5	0	0.5	0	0.5	0	0	0	1	0	**2.5**
T2	Align the corpus using some of the partners aligners in addition to the Giza++ aligner		0	0	0	0	0		0	3.5	1	1	**5.5**
T3	Collect a new and huge monolingual corpus for Arabic and English to train the language models		0	0	0		0.5	0	0	0	0	0	**0.5**
T4	Consider the possibility to use new features within Moses like exploitation of morphological analyzed corpus (alignment of Pos):		0	1	1	0.5	0.5		2	1	2	0	**8**
T5	Run Moses/partners systems with the new datasets	0	0	0	0.5	1	0		0	0	0	2	**3.5**
T6	Collect a small evaluation corpus and have it translated 4 times (or exploit existing LRs like CESTA Corpus if these were not used during phase 1)	0.5	0.5	0		0	0		0	0	0	0	**1.0**
T7	Evaluate the whole system using BLEU (and other automatic metrics) and Human Evaluations	0.5	0.5	0		0	0		0.5	0	0	1	**2.5**
PM	**Manpower/task**	**1**	**1.5**	**1**	**2**	**1.5**	**1.5**	**0**	**2.5**	**4.5**	**4**	**4**	**23.5**

10.1.7.Planning of Phase 2

Phase 1 is planned to last 12 months (January 2010-June 2010).

	Duration in months																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Phase 2																								1	2	3	4	5	6	
Phase 2-T1																														
Phase 2-T2																														
Phase 2-T3																														
Phase 2 – T4																														
Phase 2- T5																														
Phase 2- T6																														
Phase 2- T7																														

10.1.8.Deliverables of Phase 2

Phase 2 of this work package will produce the D5.3 that elaborate on the Evaluation methodology and results with recommendations for the future and delivered at T29.

Phase 2 will also update deliverable D5.2 with the new releases (if any) of the SMT developed within Phase 1 and the components added by the partners. The Language resource kit to train and evaluate the system will be also updated with the datasets of this phase.

11. Conclusions

The objective of this task within MEDAR is to end up with a baseline system exploiting existing open source toolkits as well as the partners' background. Such system will be made available. Part of the necessary Language Resources will be specified and produced.

In order to assess its performance, an evaluation platform will be set up and used. The necessary evaluation resources and tools will be produced and made available.

The evaluation platform will be made available to other players willing to evaluate their own systems in comparison to the MEDAR baseline.