Deliverable D 6.1

Machine Translation in ATLAS

Grant Agreement number: 250467
Project acronym: ATLAS
Project title: Applied Technology for Language-Aided CMS

Project type: ☐ **Pilot A** ■ **Pilot B** ☐ **TN** ☐ **BPN**

Deliverable D 6.1 Machine Translation in ATLAS

Project coordinator name, title and organisation:
Anelia Belogay, CEO, Diman Karagiozov, CTO,
Tetracom Interactive Solutions
Tel: **+35924950444**
Fax: **+35924950443**
E-mail:**anelia@tetracom.com, diman@tetracom.com**
Project website address**: www.atlasproject.eu**

Authors**:** **Cristina Vertan, University of Hamburg**
D. Karagiozov, Tetracom Interactive Solutions

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision	Date	Author, organisation	Description
0.1	08/2012	Cristina Vertan, Monica Gavrila, UHH	First Draft
0.2	09/2012	Cristina Vertan, UHH	Second Draft
0.3	11/2012	Cristina Vertan, UHH	Added Information about Domain Corpora from Svetla Koeva IBL
0.4	02/2013	Cristina Vertan, Mirela Duma, UHH	Final version , added evaluation
1.0	02/2013	Diman Karagiozov	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1 Overview of the deliverable

In this deliverable we describe the methodology used for the implementation of a machine translation (MT) engine within the ATLAS –content management system. We present first the state-of-the art of machine translation followed by a list of challenges which occur when designing a MT-engine for a content management system (Section 3). In order to select the best approach we performed a series of experiments which are described in section 3. Section 4 is dedicated to our approach for domain adaptation

2 Machine Translation – state-of-the-art overview

After more than 60 years and major progress in the development of computers and of Natural Language Processing (NLP) applications, no fully-automatic MT system has been developed, which can translate any type of input correctly. While it was initially considered a solution, word-for-word translation can only render acceptable results for the translation of some „very simple" sentences and for specific language-pairs. An MT system faces several challenges in order to obtain good translation results. These challenges may differ depending on the language-pair used: for example, while it is difficult to find word boundaries in languages like Chinese or Japanese, in European languages the word boundary is clearly represented by the 'space' character. Researchers split these problems into two categories: "linguistic" and "operational" challenges. The main linguistic challenges are ambiguity (lexical, structural, semantic etc.), text generation (lexical selection, tense generation etc.) and the mappings between the source language (SL) and the target language (TL) representations (divergences: thematic, head-switching, structural etc.): [Dorr et al., 1999] and [Somers, 2000b]. More details and examples of linguistic challenges are also presented in [Eynde,1993], [Schwarzl, 2001] and [Hutchins and Somers, 1992]. A non-exhaustive list of the operational challenges includes system maintenance, system integration with other programs and system extendibility to other domains and language pairs.

The classification of MT systems has been done according to several criteria, such as:

1. **Degree of automation**. The degree of automation is given by the amount of the user's involvement during the translation process, in this case the involvement of the human translator. Less user involvement means more system automation. Considering the degree of the user's involvement in a descending way, MT systems can be classified into three groups: Machine-aided human translation (MAHT), Human-aided MT (HAMT) and Fully automatic MT (FAMT).

2. **Type of the core technology (the paradigm)**. Regarding the core technology, the MT systems can be divided into two classes: rule-based and corpus-based (empirical). The first are often (linguistic-)theory-driven, the latter do not address either linguistic or cognitive issues. The following two MT approaches are included in the corpus-based class: statistical machine translation (SMT) and example-based machine translation (EBMT). Over the last few years, hybrid technologies have been used more frequently.

3. **Input type**. Usually an MT system has as input a text which is expected to be syntactically and semantically correct. In the last few years, systems with speech input have been developed, such as Verbmobil [Wahlster, 2000] and EuTrans-I [Amengual et al., 2000]. The translation task becomes even more complicated for speech input, as the system needs to deal also with ill-formed input. The incorrect input appears due to speech recognition errors, ungrammatical utterances etc. Incorrect text input can also evolve when, for instance, translating the output of another automatic NLP application.

4. **Level of analysis (the architecture)**. The current rule-based MT (RBMT) architectures can be organized into three classes according to the level of analysis: direct, transfer and interlingua. The first supposes a word-for-word translation from the SL to the TL, with no deeper analysis of the input than the one of the word surface forms, and with no other linguistic resources, with the exception of a bilingual dictionary. The second involves a deeper (syntactic and/or semantic) analysis and transfer rules between the SL and TL. The topmost architecture performs the translation using an intermediate (human-created) representation, which is called interlingua. Interlingua is a less ambiguous conceptual representation. Systems which use interlingua are also known as knowledge-based MT (KBMT) systems. They suppose a complete semantic representation of the input.

5. **Output quality**. The goal of MT has an impact on the expectations for translation quality. The output needs to be of high quality in MT for dissemination purposes. A comprehensible raw translation might be enough in MT for assimilation. This translation can later be edited by a human translator. A tree diagram of the MT classification according to the output quality is shown in [Carbonell et al., 1992]. A higher quality output is usually obtained when the translation domain is restricted, as in the METEO system [Chandioux, 1976].

The above classification might not be complete, as an exhaustive MT classification is beyond the scope of this document. An extended discussion on MT classification can be found in [Och, 2002] and [Schwarzl, 2001].

According to these classification criteria the system(s) we developed in this project are:

- *hybrid* system *based on combination of two corpus-based MT engines, which have as input (syntactically and semantically) well-formed text data;*

- *can be used for* translation *for assimilation.*

In the following subchapters we present a shot overview of the two approaches used: statistical machine translation and example-based machine translation.

2.1 Statistical Machine Translation (SMT)

The SMT approach has contributed to the significant resurgence in interest in MT over the last two decades. At present, there are several SMT approaches (such as word-based or phrase-based SMT) and it is by far the most widely studied MT method.

In SMT the translation process is performed by using two models: a translation model and a language model. SMT treats translation as a machine-learning problem. Formally, SMT can be defined as finding the most likely TL sentence for some SL sentence *sl*:

$$\tilde{tl} = argmax_{tl} P(sl|tl)P(tl)$$

where *tl* is a target language sentence.

An SMT system has three major components (see Osborne [2010]):

1. A translation model (TM), P (sl|tl), which specifies the set of possible translations for a source sentence and assigns probabilities to these translations. The process of extracting the TM uses a bilingual parallel aligned corpus.

2. A language model (LM), P (tl), which models the proposed target sentence. In order to obtain an LM, a monolingual corpus for the target language is needed. LMs are usually smoothed n-gram models. Usually the probability of the current word is predicted by conditioning it on two (or more) previous words.

3. A search process (the argmax operator), which is navigating through the space of possible TL translations. This process is called decoding. As this process is NP-hard for SMT, most approaches use a beam-search algorithm.

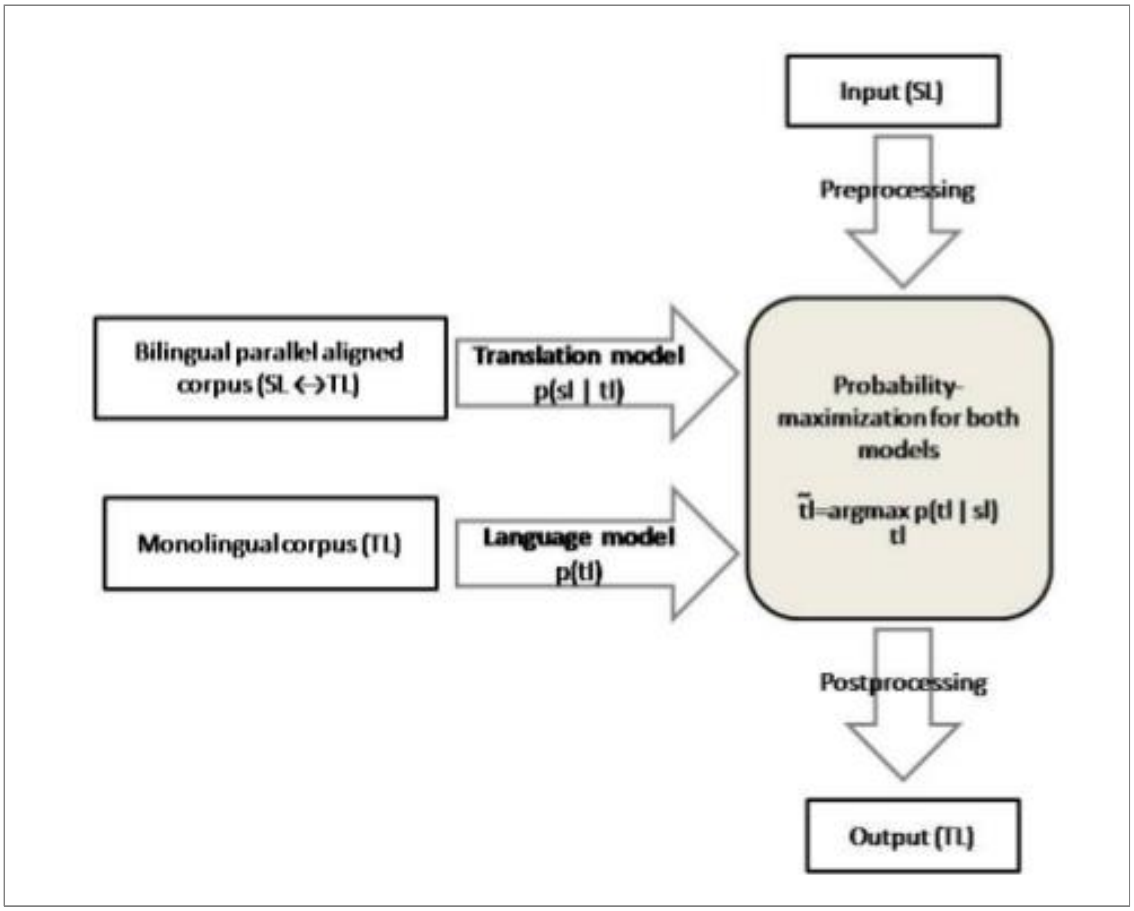

Figure 1: The SMT processes – source: [Koehn and Callison-Burch, 2005].

The SMT work-flow is shown in Figure 1. An optimal translation is obtained by maximizing the probabilities from the two models.

According to the SMT approach used (such as word-based translation like the initial IBM models [Brown et al., 1990], [Brown et al., 1993] or phrase-based translation as in [Koehn et al., 2003], [Och and Ney, 2003]) the complexity level of the models changes. A survey on SMT approaches and models is presented in [Lopez, 2008].

SMT systems can be built fast and fully automatically, provided that the needed parallel aligned corpus exists. Open-source projects, such as the phrase-based SMT system Moses (http://www.statmt.org/moses/), and the Workshop on statistical machine translation, which has been organized annually since 2006, have stimulated the development of this approach.

The quality of an SMT-output is dependent on the setting of several parameters directly for the SMT engine but also for the preprocessing steps: sentence alignment, word-to word alignment. For the ATLAS engine we performed several experiments in order to detect the optimal values. These will be presented in section 4.

2.2 Example based machine translation (EBMT)

The idea of example-based machine translation (EBMT) was first put forward in Makoto Nagao's work "A Framework of a Mechanical Translation between Japanese and English by Analogy principle" in the early 1980s [Nagao, 1984]. Since then, there has been an enormous interest in approaches which use a bilingual collection of examples (bilingual parallel aligned corpus) as the main bilingual knowledge source.

In EBMT a set of phrases in the SL and their corresponding translations in the TL are given: **the example database**. The MT system uses these examples to translate new similar SL phrases into the TL. The basic premise is that, if a previously translated phrase occurs again, the same translation is likely to be correct again. The way in which an EBMT system determines if an example is equivalent or at least similar enough to the text to be translated varies according to the approach taken by the system in creating the example database: strings, (annotated) tree structures, generalized examples (templates) etc.

After building a database of aligned examples, the "traditional" EBMT system follows three steps:

1. Matching the SL input against the example database,
2. Selecting the corresponding fragments in the TL (alignment or adaptation), and
3. Recombining the TL fragments to form a correct text (recombination). This step sometimes appears in the literature as "target sentence generation" [Kit et al., 2002] or as "synthesis" [Hutchins, 2005a].

EBMT has the big advantage of requiring less resources as the SMT. This makes it very attractive for less resourced language pairs and particular domains. Witin the ATLAS system we developed an approach which uses only the surface form of input and small databases of translation examples.

2.3 Hybrid approaches

During the last years hybrid approaches (combination of different paradigms) became very attractive, as they try to exploit benefits of each approach. The degree of hybridization varies from pipelining of engines up to full melting of two paradigms.

In the ATLAS system we choose the pipelining of an SMT and EBMT engine. The reason for this is, that especially in technical domains patterns are quite frequent; therefore there is a high chance that the input sentence is found in the translation database.

3 Requirements and challenges for the integration of Machine Translation into the ATLAS system

The integration of an MT engine into a web based content management system in general and the ATLAS system in particular, presents from the user point of view two main challenges:

1. The user may retrieve documents from different domains. Domain adaptability is a major issue in machine translation, and in particular in corpus-based methods. Poor lexical coverage and false disambiguation are the main issues when translating documents out of the training domain.

2. The user may retrieve documents from various time periods. As language changes over time, language technology tools developed for the modern languages do not work, or perform with higher error rate on diachronic documents.

With the current available technology it is not possible to provide a translation system which is domain and language variation independent and works for a couple of heterogeneous language pairs. Therefore our approach envisages a system of user guidance, so that the availability and the foreseen system-performance are transparent at any time.

The design of the system was preceded by a study of portability of results among domains and discourse genres. Especially the latter aspect plays a major role within EuDocLib where documents relevant to the European Union have to be processed. This involves: parliamentary speeches, law or news and normal regulation. As for most part of the involved languages JRC-Acquis is the only available parallel corpora within the law domain, we investigated to which extent documents within same domain but with different discourse structure can be processed by the translation engine.

Machine translation (MT) is a key component of the ATLAS WCMS, and it is embedded in all three services of the system. The development of the engine is particular challenging as the translation should be used in different domains and on different text-genres. Additionally the considered language-pairs belong most of them to the less resourced group, for which bilingual training and test material is available in limited amount.

The machine translation engine is integrated in two distinct ways into the ATLAS platform:

1. for i-Publisher (meta service for generating web sites) the MT is serving as a translation aid tool for publishing multilingual content. Text is submitted to the translation engine and the result is subject to the human post processing.

2. for i-Librarian (on-line personal digital library service, generated with i-Publisher) the MT-engine provides a translation for assimilation, which means that the user retrieving documents in different languages will use the engine in order to get a clue about the documents, and decide if he will store them for the translation is considered as acceptable it will be stored into a database.

Given the fact that the ATLAS platform deals with languages from different language families, and that the engine should support at least several domains an interlingua approach is not suitable. Building transfer systems for all language pairs is also time consuming and does not make the platform easily portable to other languages. Given the user and system requirements corpus based MT-paradigms are the only ones to be considered.

4 ATLAS Machine Translation Engine

For the MT-Engine of the ATLAS -System we decided on a hybrid architecture combining EBMT (Gavrila, 2011) and SMT at word-based level (no syntactic trees will be used) Koehn et al., 2007)). For the SMT-component part-of-speech (PoS) and domain factored models as in (Niehues and Waibel, 2010) are used, in order to ensure domain adaptability. An original approach of our system is the interaction of the MT-engine with other modules of the system:

- The document categorization module assigns to each document one or more domains. For each domain the system administrator has the possibility to store information regarding the availability of a correspondent specific training corpus. If no specific trained model for the respective domain exists, the user is provided with a warning, telling that the translation may be inadequate with respect to the lexical coverage.

- The output of the summarization module is processed in such way that ellipses and anaphora are omitted, and lexical material is adapted to the training corpus.

- The information extraction module is providing information about meta-data of the document including publication age. For documents previous to certain threshold (e.g. 1900) we do not provide translation, explaining the user that in absence of a training corpus the translation may be misleading.

The domain and dating restrictions can be changed at any time by the system administrator when an adequate training model is provided.

4.1 Architectural approach

As we mentioned in the previous section we developed a hybrid architecture pipelining an in-house EBMT engine with a Moses-based SMT-System. The architecture of the engine can be found in figure 2.

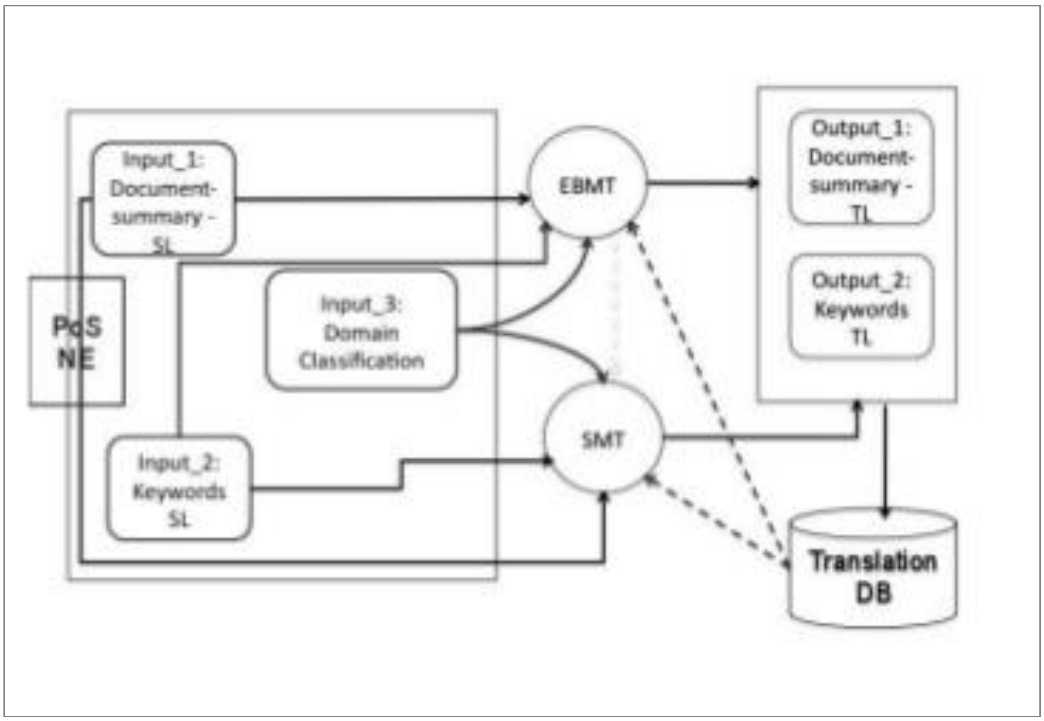

Figure 2. ATLAS MT-engine

Each of the tools depicted in the diagram are explained below.

4.1.1 Moses

Moses (http://www.statmt.org/moses) is an SMT system that enables the user to automatically train translation models for a language pair, considering that the user has the required parallel aligned corpus. The development of Moses is mainly supported under the EuroMatrix, LetsMT and EuroMatrixPlus projects, funded by the European Commission under Framework Programme 6 and 7. It received additional support from the DARPA GALE and TC-Star projects and from several universities. The tool is licensed under the GNU Lesser General Public License (LGPL).

Among the features encountered in Moses, there are:

- phrase- and tree-based translation models,
- factored translation models, which allow the integration of linguistic and other information at the word level, and
- the decoding of confusion networks and word lattices, which enable easy integration with ambiguous upstream tools, such as automatic speech recognizers or morphological analyzers.

More information about Moses can be found in [Koehn et al., 2007].

Our Moses-based MT system follows the description and the parameter setting of the baseline architecture given for the EACL 2011 Sixth Workshop on SMT. The exact parameters and training and testing steps can be found on the website of the workshop: http://www.statmt.org/wmt11/baseline.html

We trained a phrase-based model which benefits from advanced features of the decoder, such as lexicalized reordering models. In the training we used SRILM for generating the language model and GIZA++ for the alignment.

Before building the translation model (TM), the training data was preprocessed. After tokenizing the sentences, they were filtered out according to a sentence length criterion (the 'cleaning' step) and lowercased. The scripts for preprocessing the data are available on the website of the workshop. In the same way, the data for the LM was tokenized and lowercased. The language model was built with SRILM, using the parameters recommended at the Workshop: "interpolate and kndiscount". The "kndiscount" uses Chen and Goodman [1996]'s modified Kneser-Ney discounting for n-grams of order n. The "interpolate" parameter causes the discounted n-gram probability estimates at the specified order n to be interpolated with lower-order estimates.

To train the TM, we ran the provided training script. We had as input the bilingual corpus in two text files: one for the SL, the other for the TL. Each line in the SL file has a corresponding line in the TL file. For the alignment we used the default heuristics given by the value "grow-diag-final-and " of the parameter "-alignment". It starts with the intersection of the two alignments and then adds additional alignment points. As previously mentioned, a reordering model for the decoder was used. By default, only a distance-based reordering model is included in the final configuration. Additional conditional reordering models may be built and they are conditioned on specified factors (in the source and target language). These learn different reordering probabilities for each phrase pair (or just the foreign phrase). The possible configurations can be found in the Moses manual [Koehn, 2010, p. 118]. We used a "msd-bidirectional-fe" model, which considers three different orientation types: monotone, swap and discontinuous. It is conditioned on both the SL and TL phrase ("fe"). The system considers the ordering of one phrase with respect to the previous one. Using the bidirectional model, also the ordering of the next phrase with respect to the current one is modeled.

4.1.2 The SRILM toolkit

The SRI Language Modeling toolkit (SRILM) has been under development in the SRI Speech Technology and Research Laboratory since 1995 [Stolcke, 2002]. SRILM is a collection of C++ libraries, executable programs and helper scripts, which supports the creation and evaluation of a variety of language model types based on n-gram statistics, as well as several related tasks, such as statistical tagging and manipulation of n-best lists and word lattices. It runs on the UNIX and Windows platforms. It is additionally applied in several fields, such as speech recognition, machine translation, tagging and segmentation and document processing.

The SRILM toolkit is freely available under an open source community license andcan be downloaded from http://www-speech.sri.com/projects/srilm. It is currently used in the research community for tasks requiring statistical language modeling.

The ATLAS engine uses the version 1.5.7 of the SRILM toolkit for creating the LM of the Moses-based MT system and extracting the necessary information for the recombination step of the implemented EBMT systems.

4.1.3 GIZA ++

GIZA++ was developed by Franz Josef Och [Och and Ney, 2003] and is an extension of the program GIZA, which was part of the SMT toolkit EGYPT23. It can be used to train the IBM Models 1-5 [Brown et al., 1993] and an HMM word alignment model [Vogel et al., 1996]. The package also contains the source for the mkcls tool which generates the word classes necessary for training some of the alignment models.

GIZA++ can be freely used under the terms of GNU Public License (GPL) version 2 and is available on http://code.google.com/p/giza-pp/24 . It is known to compile on Linux, Irix and SUNOS systems.

The version we used in ATLAS system is 1.0.2. We needed GIZA++ to run the Moses-based SMT system and to obtain the word-alignments in the EBMT system(s)

4.1.4 EBMT engine

Lin-EBMT is a linear EBMT system, which in the recombination step makes use only of the word sequences provided by the alignment. The output is formed by employing only the 2-gram information extracted from the corpus and the recombination matrix. From the matching step, not only the SL sentences which best cover the input are obtained, but also the corresponding TL sentences.

4.1.4.1 Matching the input

In the matching step, the input sentence is compared with the sentences extracted from the corpus after using the index. The algorithm tries to match the (whole) input with an entry in the corpus. In cases where this is not possible, it tries to match parts of the input with (parts of) the sentences in the corpus.

Require: the input I, the ids of the sentences which have at least one token in common with the input (punctuation excluded): $\{sentenceId_1, ..., sentenceId_m\}$.
Ensure: the set of the M matched sentences together with the corresponding longest common subsequences: $Result = \{(sentenceId_i, LCS_i)\}$, $1 \leq i \leq M$, where $sentenceId_i$ is the id of the matched sentence S_i and LCS_i is the longest common subsequence between I and S_i.
$Result = \emptyset$
$copyI \leftarrow I$
$IDS \leftarrow \{sentenceId_1, ..., sentenceId_m\}$
while $copyI \neq \epsilon$ **do**
{ϵ is the empty string.}
$(id, LCS) \leftarrow findBestMatch(copyI, IDS)$
$Result \leftarrow Result \cup \{(id, LCS)\}$
$copyI \leftarrow remove(copyI, LCS)$
$IDS \leftarrow IDS \setminus \{id\}$
end while

Figure 3: Matching algorithm in EBMT

4.1.4.2 Alignment

The required word alignment information is extracted at run-time from the GIZA++ output obtained while running the Moses-based SMT system. From the two generated 'A3.final ' files, only the target-

source language direction file is consulted for the implementation in this thesis. The 'A3.final ' file contains the final word-to-word alignment for each of the words in each line (in the same order as the input parallel aligned corpus).

4.1.4.3 Recombination

The recombination algorithm is based on the monolingual distribution of bi-grams and on a "recombination matrix "

$$A_{N,N} = (a_{i,j})_{1 \le i,j \le N} = \begin{cases} -3, & \text{if } i = j; \\ -2, & \text{if } i \neq j, \\ & count(w_{i_{last}} w_{j_1}) = 0; \\ \frac{2 * count(w_{i_{last}} w_{j_1})}{count(w_{i_{last}}) + count(w_{j_1})}, & \text{if } i \neq j, \\ & count(w_{i_{last}} w_{j_1}) \neq 0. \end{cases}$$

where $count(s)$ represents the number of appearances of a token s in the corpus.

More details about the algorithm can be found in [Gavrila 2012].

4.2 Corpora used for the evaluation and validation

4.2.1 Acquis Communautaire

The **Acquis Communautaire (AC)** (http://wt.jrc.it/lt/Acquis)is the total body of European Union law applicable in the EU Member States. This collection of legislative texts contains texts written between the 1950s and today and it changes continuously. The texts are available in all official EU languages but Irish, i.e. in 22 languages. The Language Technology group of the European Commission's Joint Research Center and the Romanian Academy of Sciences processed, aligned and encoded part of these texts and created the JRC-Acquis corpus, which is seen as "an approximation of the Acquis Communautaire".

The corpus consists of around 20 000 documents with an average of 47 million words per language. It is XML encoded, following the Text Encoding Initiative Guidelines. In the table below we illustrate the difference in number of words sentences across the monolingual parts in JRC-Acquis for 3 languages: German, Romanian and English.

Language	**No. texts**	**No. words** (Text body)	**No. words** (Signatures)	**No. words** (Annexes)	**Total no. words** (Whole document)
German	23541	32059892	2542149	16327611	50929652
English	23545	34588383	3198766	17750761	55537910
Romanian (version 1)	6573	9186947	514296	11185842	20887085
Romanian (version 2)	19211	30832212	-	-	30832212

Table1. Statistics for the monolingual corpora in JRC-Acquis (Romanian, English and German).

Table 2 shows the influence on the different size on monolingual corpora on the parallel alignments. Thus, even within JRC-Acquis we have different size of training data across different language pairs.

Language pair	No. of documents	No. of links
German-Romanian	6558 docs	391972 links
German-English	23430 docs	1264043 links
English-Romanian	6557 docs	391334 links

Table 2. Statistics on bilingual alignments in JRC-Acquis

4.2.2 ROGER

RoGER is a parallel corpus, aligned at sentence level. It is domain-restricted, as the texts are from a users' manual of an electronic device. The languages included in the development of this corpus are Romanian, English, German and Russian. The corpus was manually compiled. It is not annotated and diacritics are ignored. The corpus was manually verified: the translations and the (sentence) alignments were manually corrected.

The initial PDF files of the manual were transformed into text (.RTF) files, where graphics and pictures were either left out (pictures around the text), or replaced with text (pictures inside the text). The initial text was preprocessed by replacing numbers, websites and images with "meta-notions" as follows: numbers by NUM, pictures by PICT and websites by WWWSITE. In order to simplify the translation process, some abbreviations were expanded. The sentences were manually aligned, first for groups of two languages. Finally, the two alignment files obtained were merged, so that, after all, RoGER contained all four languages. The merged text files are XML encoded.

The corpus contains 2333 sentences for each language. More statistical data about the corpus is presented in Table 3. The average sentence length is eleven tokens for English, Romanian and German and nine for Russian. The tokens can be a lexical item, a punctuation sign or a number. More about the RoGER corpus can be found in [Gavrila and Elita,2006]

Feature	English	Romanian	German	Russian
No. tokens	26096	25850	27142	22383
Vocabulary size	2012	3104	3031	3883
Vocabulary (*Word-frequency higher than two*)	1231	1575	1698	1904

Table 3. ROGER statistics

4.2.3 Domain corpora

In order to perform domain adaptation we compiled small parallel corpora for all 13 domains included in the ATLAS initial hierarchy. These corpora contain at least 2000 sentences. The number of 2000 sentences is based on experiments performed in [Gavrila 2012].

The approach was the following:

1. Fill as many categories as possible with original parallel data from several main sources (European Central Bank corpus, SETimes, BulNC, etc.) and other texts from the Internet. No limit is placed on original parallel data and all available texts are used.

Lang. pair	Full	Partially filled	Empty	Domain	Full	Partially filled	Empty
EN-BG	6	4	3	Business	15	0	0
EN-DE	6	4	3	Computing	9	3	2
EN-EL	4	3	6	Fiction	6	2	7
EN-PL	6	3	4	History	0	15	0
EN-RO	6	3	4	Politics	15	0	0
BG-DE	4	4	5	Maths	0	0	15
BG-EL	4	4	5	Physics	15	0	0
BG-PL	6	1	6	Biology	0	0	15
BG-RO	5	3	5	Chemistry	0	0	15
DE-EL	5	1	7	Arts	0	8	7
BG-PL	5	1	7	Music	0	7	8
BG-RO	5	2	6	Geography	1	5	9
EL-PL	5	1	7	Sociology	15	0	0
EL-RO	4	3	6				
PL-RO	6	1	6				

Table 4 . Availability of original parallel texts. For each language pair, the number of filled, partialand empty domain corpora (cols. 1-4). For each domain, the number of filled, partial and empty language pairs (cols. 5-8)

2. If a domain is not filled for a particular pair of languages L1_L2, all unused original texts in either L1 or L2 are collected and used for MT. Such files can be found in other folders containing L1 or L2 in their names E.g., if we have the following contents of the corresponding subfolders:

```
EN_PL: file1en.txt, file1pl.txt, file2en.txt, file2pl.txt
EN_RO: file1en.txt, file1ro.txt, file3en.txt, file3ro.txt,
```

we can use file file2en.txt to generate Romanian translation and file3en.txt to generate Polish translation but we cannot use file1en.txt in neither case as it is already present in both folders.

3. Collect monolingual texts and generate parallel corpora via MT for unfilled categories – as a result **all 195 corpora have been filled** (methodology for generating parallel texts is explained below).

Generation of parallel corpora using MT

1) Stages:
   a) Automatic sentence segmentation of the SL
   b) Selection of appropriate sentences for translation
   c) Obtaining MT from the source
   d) Evaluating and selecting parallel pairs
2) Different approaches have been examined for selection of suitable sentences for MT considering length (80-800 symbols), punctuation (end of sentence) and contents (at least 70% letters)
3) MT systems used:
   a) Google Translate. Free access with extended daily limit provided via University Research Program.
   b) WebTrans by SkyCode. Free access was provided by SkyCode.
   c) Bing (Microsoft Translator). Limited access – 2 million symbols per month.
   d) MyMemory (translation memory). Free access.
4) Several approaches for selection of proper parallel sentences after MT: full or partial (>50% or >70%) overlap of translations from two MT systems – based either on string, tokens or lemmas; BLEU between translations from two MT systems (>0.4 or >0.7); BLEU score measured between the SL text and back translation of MT.

4.3 Domain adaptation

Having a parallel corpora consisting of aligned data from the source language and from the target language, SMT can be performed and thus, a system that is able to translate from the source language to the target language is created. But, semantics is an important aspect that has to be considered. Usually, the corpora used in the training of the system are domain-specific (out-of-domain). When using and testing the system on a different domain (in-domain), the system can give very poor results at evaluation even though it had good results when it was evaluated on out-of-domain data. This is due to the fact that words and phrases have different meanings in different domains and the vocabulary, style and even grammar can differ from one domain to another. The task of creating a translation system that is able to give the expected result when input from different domains is given is referred to as domain adaptation. As noted in (Bellegarda, 2004), the high

variability of natural language confirms the need of adaptation. Bellegarda points out that the evolution of language is a direct consequence of the dynamic time and world we live in. New terms (technical, domain-specific terms) are added to the lexicon every day. Moreover, the same concept can be interpreted differently depending on the domain. Also, a text (or speech) can be written in a technical style or in an informal style. Different domains can have different discourse styles that can be influenced by the emotional state of a person. This variability of syntactic and semantic features and the dynamic change of language lead to challenges in the task of domain adaptation.

In this section the state-of-the-art for domain adaptation for statistical machine translation is presented. DA could be classified by taking into consideration the sub-tasks (sub-processes) that occur in building a SMT system: alignment model, language model, translation model, reordering model. In the next section, related work that deals with the adaptation of one or more of these models is described.

4.3.1 Alignment Model adaptation

(Ker & Chang, 1997) developed an algorithm that identifies words and their "in-context" translations in a bilingual corpus. They used a class-based method in order to align the words and have gained good results, especially for precision.

(Och & Ney, 2000) presented in their paper different extensions to statistical alignment models. It is proved that the models IMB-1 and IMB-2 perform worse than alignment models with a first-order dependence.

(Wu, Wang, & Liu, 2005) performed experiments in alignment adaptation. They used out-of-domain data in order to get better results at in-domain word alignment. In their work, an alignment model is trained using the out-of-domain corpus and another alignment model is trained using the in-domain corpus (out-of-domain >> in-domain). By interpolating the two models, a new alignment model results and also by interpolating the two dictionaries that correspond to the domains, a new dictionary is created. The dictionary is used in improving the adaptation results. The proposed method achieved improvements in both precision and recall.

4.3.2 Language Model adaptation

(Zhao, Ji, Xi, Huang, & Chen, 2011) worked on language model adaptation focusing on weight modeling. In order to measure the similarity between different corpora, the cross-entropy of translation output is used as a metric. Even though only the language model weight is tuned, this adaptation technique performs much better than a baseline.

In (Bulyko, Matsoukas, Schwartz, Nguyen, & Makhoul, 2007) the adaptation method is based on optimizing the language model mixture weights using TER or BLEU. Using n-best lists generated with a development set and tuning the weights using Powell's hill climbing algorithm, TER is minimized on the set of n-best lists (BLEU is maximized). In their experiments, Bulyko et al. analyzed two methods of combining language model components: log-linear combined probabilities and language model interpolation. The results show a maximum improvement for translation of 0.4 BLEU and a reduced translation edit rate of 0.2%, respectively.

Language model interpolation is also investigated by (Koehn & Schroeder, 2007). For language model interpolation, the SRILM toolkit is used. First, a language model for the out-of-domain corpus and a

language model for the in-domain are obtained. Then, the best weight is computed using the SRILM toolkit and then the two language models are interpolated. However, this method had a slightly lower BLEU score of 27.12 compared to using only the language model trained on the in-domain (BLEU of 27.46).

4.3.3 Translation Model adaptation

Regarding translation model adaptation, (Koehn & Schroeder, 2007) used in their experiments the combined training data to obtain a translation model and also two translation models using the factored translation model framework. The method makes use of alternative decoding paths, available in Moses toolkit. Two decoding paths were used: the in-domain translation table and the out-of-domain translation table. Among all their experiments from the paper, the alternative decoding path method gained the best BLEU score.

(Callison-Burch, Koehn, & Osborne, 2006) focus on improving the translation using paraphrasing. Their work addresses the problem of unknown phrases. When an unknown source phrase is encountered, a paraphrase derived from external resources to the training corpus can be used in the translation. The method implies using paraphrase probabilities defined using two probabilities: the probability that the initial phrase in English can be translated as a particular phrase in another language and the probability that the candidate paraphrase translates as the foreign language phrase. Their results showed improvements in the BLEU score even though the authors argue that BLEU is not the best metric that can be used to evaluate the method of using paraphrasing as BLEU uses exact matches of n-grams in a reference translation.

Another approach to translation model adaption relies on using comparable corpora. In (Snover, Dorr, & Schwartz, 2008), monolingual target data is used in the improvement of an SMT system. The method consists in using multiple texts in the target language that have similar topic as the source language document that will be translated. Documents that could have similar topic to the source documents are searched in a monolingual corpus in the target language. Then the documents are used to increase the probability of generating texts that are similar to the comparable document. By using this method, the BLEU score obtained is higher than the baseline system score.

(Bertoldi & Federico, 2004) used monolingual resources to perform domain adaptation. Their method makes use of either source or target large monolingual in-domain data. A bilingual corpus is created by translating the monolingual adaptation data into the other language and then is used to adapt the translation and the reordering model. Also, another approach used in this work is adapting the language model using synthetic or given text in the target language. The performance of the system improves over the baseline system when using language model, translation model and reordering model adaptation, especially when language model adaptation is used.

A technique that weights automatically the corpora and alignments is described in (Shah, Barrault, & Schweng, 2010). The method is referred to as resampling. A number of parallel corpora are concatenated and word alignments are extracted using GIZA++. The alignments are separated depending on the parallel corpora it used and weights are computed for each alignment. An algorithm for resampling the alignments is presented and also an algorithm for weighting the corpora. When using this method, the BLEU score improved.

4.3.4 Reordering Model adaptation

(Ling, Luis, Graca, Coheur, & Trancoso, 2011) use weighted alignment matrices for reordering modeling. These matrices encode all possible alignments and generate better phrase-tables. The alignment matrix is used to create the translation model and the 1-best alignment to generate the reordering model. In their paper, two algorithms to generate the reordering model are presented: one uses the alignments for the phrase pairs, and the other algorithm makes use of the contextual information of the phrase pairs. Increased BLEU score is obtained at evaluation.

In (Chen, Zhang, Aw, & Li, 2008) n-best hypotheses are used for language, translation and reordering model adaptation. Each hypotheses holds phrase alignment information that is useful in the word reordering for the source text. The best word reordering for a source text is the one with the highest posterior probability. The source sentences are reordered taking into consideration the best word reordering. The weights of the decoder are optimized using the reordered source sentences. Using this method, the BLEU score was improved over the baseline system.

5 Evaluation

The evaluation of machine translation (MT) systems is still an open research issue. There are two main directions in the evaluation of MT-systems:

1. the human evaluation following criteria like intelligibility, readability, time used for correction
2. the automatic evaluation based on measures like BLEU, TER, METEOR

Given the number and the diversity of language pairs covered by the ATLAS system, a manual evaluation was impossible.

The ATLAS-MT-engine contains domain specific models for 13 domains. The adaptation was done by injecting a small in-domain parallel corpus into the larger and more general JRC-Acquis corpus. For evaluation purposes, we isolated before training **5% from the in-domain corpus**, as test data. The test-data was not involved in the training process.

As the ATLAS-engine is a corpus-based approach we were interested to see how broad the language coverage is. We decided to measure as indicator the number of un-translated words in the test-set mentioned above.

5.1 BLEU

BLEU (bilingual evaluation understudy), one of the evaluation scores applied most frequently for MT evaluation, measures the number of n-grams of different lengths of the system output that appear in a set of references. More details about BLEU can be found in [Papineni et al., 2002].

Although criticized more recently, it is still important to calculate the BLEU score for comparison reasons, as for many previous developed systems it is the only evaluation measure available. The BLEU score is computed according to the following formula:

$$BLEU = BP * exp(\sum_{n=1}^{N} \frac{1}{N} log(p_n))$$

where N is the maximum n-gram size and the brevity penalty BP is calculated as:

$$BP = min(1, e^{1-\frac{r}{c}})$$

c is the length of the corpus of hypothesis translations and r is the effective reference corpus length. The value for r is calculated as the sum of the single reference translation from the each set which is closest to the hypothesis translation.

Papineni et al. [2002] calculate the n-gram precision pn as the sum over the matches for every hypothesis sentence S in the complete corpus C as:

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} count(ngram)}$$

For the evaluation with BLEU, we used the twelfth version of the NIST/BLEU im-plementation provided by www.itl.nist.gov.

5.2 Evaluation results

The evaluation of machine translation engine cannot be done overall but per domain, as we have separate translation models per domain. The size of the in-domain training data varies from one domain to another. Overall the ATLAS-MT systems relies mainly on 390 translation subsystems (2 models per language pair x 15 language pairs x 13 domains)

In the following tables we present the results for MT_BLEU for two domains, one of scientific nature - Computer Science and one related to humanities: Politics.

5.2.1 POLITICS

Language Pair	BLEU score	#sentences in-domain corpus	#sentences out-domain corpus	#sentences test set
BG-DE	8.70	1914	344997	100
BG-EL	28.14	52930	515072	2750
BG-RO	16.53	59371	241670	3000
BG-EN	39.36	56796	306767	3000
BG-PL	7.05	1981	367523	100
RO-DE	8.88	2261	324448	100
RO-EL	4.41	53613	159417	2750

RO-BG	19.09	59371	241670	3000
RO-EN	39.12	64329	336455	3351
*RO-PL	6.12	1912	362321	100
EL-RO	13.91	53613	159417	2750
EL-EN	36.68	51564	592923	2750
EL-BG	31.76	52930	515072	2750
EL-PL	5.54	1963	641689	100
EL-DE	12.01	1912	719960	100
*PL-RO	9.76	1914	362321	100
PL-EN	14.15	1889	1183516	100
PL-DE	4.00	1930	1179492	100
PL-EL	4.45	1963	641689	100
PL-BG	14.62	1981	367523	100
EN-DE	19.75	1988	1199447	100
EN-EL	30.81	51564	592923	2750
EN-RO	39.75	64329	336455	3351
EN-BG	0.31	56796	306767	3000
EN-PL	8.97	1889	1183516	100
DE-EN	24.35	1988	1199447	100
DE-PL	13.72	1930	1179492	100
DE-RO	4.64	2261	324448	100
DE-EL	10.82	1912	719960	100
DE-BG	13.32	1914	344997	100

Table 5. BLEU score for domain Politics

5.2.2 Business

Language Pair	BLEU score	#sentences in-domain corpus	#sentences out-domain corpus	#sentences test set
BG-DE	12.94	2037	344997	100
BG-EL	21.20	9773	515072	500
BG-RO	11.33	10410	241670	500
BG-EN	31.88	9321	306767	500

BG-PL	12.56	2063	367523	100
RO-DE	10.41	1918	324448	100
RO-EL	17.52	9774	159417	500
RO-BG	19.58	10410	241670	500
RO-EN	23.82	10109	336455	500
RO-PL	12.28	1921	362321	100
EL-RO	11.58	9774	159417	500
EL-EN	22.26	83371	592923	4300
EL-BG	24.08	9773	515072	500
EL-PL	5.12	25853	641689	1350
EL-DE	12.08		719960	
PL-RO	11.98	1921	362321	100
PL-EN	7.78	62838	1183516	3000
PL-DE	1.39	39729	1179492	2000
PL-EL	5.42	25853	641689	1350
PL-BG	13.07	2063	367523	100
EN-DE	11.30	93160	1199447	4500
EN-EL	18.03	83371	592923	4300
EN-RO	14.97	10109	336455	500
EN-BG	27.98	9321	306767	500
EN-PL	6.31	62838	1183516	3000
DE-EN	13.18	93160	1199447	4500
DE-PL	1.03	39729	1179492	2000
DE-RO	7.92	1918	324448	100
DE-EL	17.04		719960	
DE-BG	12.71	2037	344997	100

Table 6. BLEU score for domain Business

5.2.3 Computing

Language Pair	BLEU score	#sentences in-domain corpus	#sentences out-domain corpus	#sentences test set
BG-DE	16.33	2536	344997	125

BG-EL	20.90	2352	515072	100
BG-RO	17.26	3061	241670	150
BG-EN	27.29	2642	306767	125
BG-PL	15.60	2085	367523	100
RO-DE	7.79	4836	324405	250
RO-EL	20.50	2046	159417	100
RO-BG	16.72	3061	241670	150
*RO-EN	5.21	5133	336455	250
RO-PL	3.41	3566	362321	175
EL-RO	21.23	2046	159417	100
*EL-EN	7.20	3963	592923	200
EL-BG	19.19	2352	515072	100
*EL-PL	3.43	1963	641689	125
EL-DE	11.53	4439	719960	255
PL-RO	1.36	3566	362321	175
*PL-EN	3.6	9013	1183516	500
*PL-DE	2.99	8384	1179492	425
*PL-EL	2.75	1963	641689	125
PL-BG	14.59	2085	367523	100
*EN-DE	30.21	40459	1199447	2125
*EN-EL	7.75	3963	592923	200
*EN-RO	4.45	5133	336455	250
EN-BG	19.12	2642	306767	125
*EN-PL	2.06	9013	1183516	500
*DE-EN	2.70	40459	1199447	2125
*DE-PL	0.85	8384	1179492	425
DE-RO	6.07	4836	324405	250
DE-EL	11.54	4439	719960	255
DE-BG	14.49	2536	344997	125

Table 7. BLEU score for domain Computing

5.3 Comparative assessment

Given the fact that corpus-based MT-systems are extremely sensible to the type and amount of training data, only an approximate comparison is possible. The very last results published (shown in the table below) by META-NET in 2012 rely on tests done on in-domain training data from JRC Acquis communautatire.

	Zielsprache – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

Table 8: BLEUGold-standard measurements as in “The German Language in the digital age”- A. Burchardt, M. egg, K. eichler, B. Krenn, J. Kreutel, A. Leßmöllmann, g. Rehn, M. stede, H. Uszkoreit, M. Volk., METa-NET White papers , Springer 2012

Table 9 summarizes the relevant gold values and the results obtained for the ATLAS MT-System

Language Pair	Computing	Politics	Business	Gold
BG-DE	16.33	8.70	12.94	38.7
BG-EL	20.90	28.14	21.20	34.5
BG-RO	17.26	16.53	11.33	36.8
BG-EN	27.29	39.36	31.88	61.3
BG-PL	15.60	7.05	12.56	35.1
RO-DE	7.79	8.88	10.41	38.5
RO-EL	20.50	4.41	17.52	35.6
RO-BG	16.72	19.09	19.58	33.1

*RO-EN	5.21	39.12	23.82	60.8
RO-PL	3.41	6.12	12.28	35.8
EL-RO	21.23	13.91	11.58	37.2
*EL-EN	7.20	36.68	22.26	59.5
EL-BG	19.19	31.76	24.08	32.4
*EL-PL	3.43	5.54	5.12	34.2
EL-DE	11.53	12.01	12.08	43.1
PL-RO	1.36	9.76	11.98	38.2
*PL-EN	3.6	14.15	7.78	60.8
*PL-DE	2.99	4.00	1.39	40.2
*PL-EL	2.75	4.45	5.42	34.2
PL-BG	14.59	14.62	13.07	31.5
*EN-DE	30.21	19.75	11.30	46.8
*EN-EL	7.75	30.81	18.03	41.0
*EN-RO	4.45	39.75	14.97	49.0
EN-BG	19.12	0.31	27.98	40.5
*EN-PL	2.06	8.97	6.31	49.2
*DE-EN	2.70	24.35	13.18	53.6
*DE-PL	0.85	13.72	1.03	30.2
DE-RO	6.07	4.64	7.92	30.7
DE-EL	11.54	10.82	17.04	32.8
DE-BG	14.49	13.32	12.71	26.3

Table 9 Gold-standard BLEU and values for 2 domains in ATLAS system

We observe that the scores obtained for the ATLAS system are below the gold scores. This is not an indication for the quality of the systems as in fact one evaluation campaign cannot be replicated. Our tests are performed on different test sets, on different domains as the Gold Standard.

The systems marked with "*" are still in development, in the sense that the sentence-alignment of the "in-domain data" has to be revised. The wrong automatic sentence alignment has a major influence on the translation quality.

6 ANNEX

6.1 Training script for domain adaptation

6.1.1 1. Prepare data

- tokenize in-domain data and out-of-domain data

~/local/experiment$ ~/local/tools/moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < ~/local/experiment/de-en_in.en > ~/local/experiment/de-en_in.tok.en

~/local/experiment$ ~/local/tools/moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l de < ~/local/experiment/de-en_in.de > ~/local/experiment/de-en_in.tok.de

~/local/experiment$ ~/local/tools/moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l en < ~/local/experiment/de-en_out.en > ~/local/experiment/de-en_out.tok.en

~/local/experiment$ ~/local/tools/moses/mosesdecoder/scripts/tokenizer/tokenizer.perl -l de < ~/local/experiment/de-en_out.de > ~/local/experiment/de-en_out.tok.de

- lowercase data from both domains

~/local/tools/moses/mosesdecoder/scripts/tokenizer/lowercase.perl < ~/local/experiment/de-en_in.tok.en > ~/local/experiment/de-en_in.lowercased.en

~/local/tools/moses/mosesdecoder/scripts/tokenizer/lowercase.perl < ~/local/experiment/de-en_in.tok.de > ~/local/experiment/de-en_in.lowercased.de

~/local/tools/moses/mosesdecoder/scripts/tokenizer/lowercase.perl < ~/local/experiment/de-en_out.tok.en > ~/local/experiment/de-en_out.lowercased.en

~/local/tools/moses/mosesdecoder/scripts/tokenizer/lowercase.perl < ~/local/experiment/de-en_out.tok.de > ~/local/experiment/de-en_out.lowercased.de

- filter out long sentences from both domains

~/local/tools/moses/mosesdecoder/scripts/training/clean-corpus-n.perl de-en_in.lowercased de en de-en_in.clean 1 80

~/local/tools/moses/mosesdecoder/scripts/training/clean-corpus-n.perl de-en_out.lowercased de en de-en_out.clean 1 80

- concatenate the source language corpus from both domains: de-en_in.clean.de + de-en_out.clean.de ==> de-en.clean.de

- concatenate the target language corpus from both domains: de-en_in.clean.en + de-en_out.clean.en ==> de-en.clean.en

6.1.2 2. Build interpolated language model

- build language model for in-domain data

~/local/tools/srilm/bin/i686/ngram-count -order 3 -interpolate -kndiscount -unk -text ~/local/experiment/de-en_in.lowercased.en -lm ~/local/experiment/lm/lm_in.lm

- build language model for out-of-domain data

~$ ~/local/tools/srilm/bin/i686/ngram-count -order 3 -interpolate -kndiscount -unk -text ~/local/experiment/de-en_out.lowercased.en -lm ~/local/experiment/lm/lm_out.lm

- use the concatenated corpus (target language) from in-domain and out-of-domain to build ppl files

~/local/tools/srilm/bin/i686/ngram -order 3 -unk -lm ~/local/experiment/lm/lm_in.lm -ppl ~/local/experiment/de-en.lowercased.en -debug 2 > ~/local/experiment/ppl_files/ppl1.ppl

~$ ~/local/tools/srilm/bin/i686/ngram -order 3 -unk -lm ~/local/experiment/lm/lm_out.lm -ppl ~/local/experiment/de-en.lowercased.en -debug 2 > ~/local/experiment/ppl_files/ppl2.ppl

- use ppl files to compute the best lambda = L

~/local/tools/srilm/bin/i686/compute-best-mix ~/local/experiment/ppl_files/ppl1.ppl ~/local/experiment/ppl_files/ppl2.ppl

- use language models and lamda to interpolate the language models

~/local/tools/srilm/lm/bin/i686/ngram -lm ~/local/experiment/lm/lm_in.lm -mix-lm ~/local/experiment/lm/lm_out.lm -lambda L -write-lm ~/local/experiment/lm/mixlm.lm -unk

6.1.3 3. Train phrase model

- use the concatenated corpora and the interpolated language model to train the phrase model

~$ nohup nice ~/local/tools/moses/mosesdecoder/scripts/training/train-model.perl -root-dir train -corpus ~/local/experiment/de-en.clean -f de -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe -lm 0:3:$HOME/local/experiment/lm/mixlm.lm:8 -external-bin-dir ~/local/tools/bin >& training.out

6.1.4 4. Example of test

~/local/experiment$ nohup nice ~/local/tools/moses/mosesdecoder/bin/moses -f ~/local/experiment/train/model/moses.ini < ~/local/experiment/test_in.de > ~/local/experiment/test_translated.en 2> ~/local/experiment/test.out

~/local/experiment$ ~/local/tools/moses/mosesdecoder/scripts/generic/multi-bleu.perl -lc ~/local/experiment/test_in.en < ~/local/experiment/test_translated.en

7 References

Juan Carlos Amengual, Jose Miguel Benedi, Francisco Casacuberta, Asuscion Castano, Antonio Castellanos, Victor Jimenez, David Llorens, Andreas Marza, Moises Pastor, Federico Prat,

Bellegarda, J. (2004). Statistical language model adaptation: review and perspectives. Speech Communication .

Bertoldi, N., & Federico, M. (2004). Domain Adaptation for Statistical Machine Translation with Monolingual Resources. Journal Information Retrieval .

Bulyko, I., Matsoukas, S., Schwartz, R., Nguyen, L., & Makhoul, J. (2007). Language Model Adaptation in Machine Translation from Speech. Acoustics, Speech and Signal Processing.

Jaime G. Carbonell, Teruko Mitamura, and Eric Nyberg. The kant perspective: A critique of pure transfer (and pure interlingua, pure statistics, . . .). In Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Empiricist vs. Rationalist methods in MT, pages 225–235., Montreal, June 1992. CCRIT-CWARC.

John Chandioux. Meteo, an operational system for the translation of public weather forecasts.

American Journal of Computational Linguistics, microfiche 46:pp.27–36, 1976.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). Improved Statistical Machine Translation Using Paraphrases. Proceeding HLT-NAACL '06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics.

Chen, B., Zhang, M., Aw, A., & Li, H. (2008). Exploiting N-best Hypotheses for SMT Self-Enhancement. Proceedings of ACL-08: HLT, Short Papers (Companion Volume).

Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A survey of current paradigms in machinetranslation. Advances in Computers, 49:2–68, 1999.

Miles Osborne. Encyclopedia of Machine Learning, chapter Statistical Machine Translation.Springer, 2010. URL http://www.statmt.org/ued/?n=Public.Publications. Editor: Claude Sammut and Geoffrey I. Webb. 12

Frank Van Eynde, editor. Linguistic Issues in Machine Translation. Pinter Publishers, London and New York, 1993.

Monica Gavrila, Improving Recombination in a Linear EBMT System by Use of Constraints, http://ediss.sub.uni-hamburg.de/volltexte/2012/5758/ , 2012

Monica Gavrila and Natalia Elita. Roger - un corpus paralel aliniat. In Resurse Lingvistice¸in Instrumente pentru Prelucrarea Limbii Romˆne Workshop Proceedings, pages 63–67, December 2006. Workshop held in November 2006, Publisher: Ed. Univ. Alexandru Ioan Cuza, ISBN:978-973-703-208-9

W. John Hutchins and Harold L. Somers. An Introduction to Machine Translation. Academic

Press, London, 1992. ISBN: 0-12-362830-X.

Ker, S. J., & Chang, J. S. (1997). A class-based approach to word alignment. Association for Computational Linguistics .

Koehn, P., & Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation.

Ling, W., Luis, T., Graca, J., Coheur, L., & Trancoso, I. (2011). Reordering Modeling usingWeighted Alignment Matrices. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.

Adam Lopez. Statistical machine translation. ACM Comput. Surv., 40(3):1–49, 2008. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/1380584.1380586.

Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In Proceedings of the international NATO symposium on Artificial and human intelligence, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc. ISBN0-444-86545

Franz Josef Och. Statistical Machine Translation: from Single-World Models to AlignmentTemplates. PhD thesis, Facultaet fuer Mathematik, Informatik und Naturwissenschaften der Rheinisch-Westfaelischen Technischen Hochschule Aachen, 2002.

Och, F. J., & Ney, H. (2000). Improved Statistical Alignment Models. Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Session: Machine translation and evaluation, pages 311 – 318, Philadelphia, Pennsylvania, 2002. Publisher: Association for Computational Linguistics Morristown, NJ, USA.

Shah, K., Barrault, L., & Schweng, H. (2010). Translation Model Adaptation by Resampling. Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR.

Anja Schwarzl. The (Im)Possibilities of Machine Translation. European University Studies: Series XIV, Angla-Saxon Language and Literature. PETER LANG, 2001

Harold Somers. Handbook of Natural Language Processing, chapter Machine Translation, pages329–346. Marcel Dekker Inc, 2000b. Editors: Robert Dale and Hermann Moisl and Harold Somers.

Snover, M., Dorr, B., & Schwartz, R. 2008. Language and Translation Model Adaptation using Comparable Corpora. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.

Wu, H., Wang, H., & Liu, Z. (2005). Alignment Model Adaptation for Domain-Specific Word Alignment. Proceedings of the 43rd Annual Meeting of the ACL.

Zhao, Y., Ji, Y., Xi, N., Huang, S., & Chen, J. (2011). Language ModelWeight Adaptation Based on Cross-entropy for Statistical Machine Translation. 25th Pacific Asia Conference on Language, Information and Computation.

Enrique Vidal, and Juan Miguel Vilar. The eutrans-i speech translation system. MachineTranslation, 15(1⁄2):75–103, 2000.

Wolfgang Wahlster, editor. Verbmobil: Foundations of Speech-to-Speech Translation. Springer Verlag, 2000