Integration of Machine Translation in On-line Multilingual Applications – Domain Adaptation

Mirela-Ştefania Duma
University of Hamburg, Germany
mduma@informatik.uni-hamburg.de

Cristina Vertan
University of Hamburg, Germany
cristina.vertan@uni-hamburg.de

Large amounts of bilingual corpora are used in the training process of statistical machine translation systems. Usually a general domain is used as the training corpus. When the system is tested using data from the same domain, the obtained results are satisfactory, but if the test set belongs to a different domain, the translation quality decreases. This is due to insufficient lexical coverage, wrong choice in case of polysemous words and differences in discourse style between the two domains. Thus, the need to adapt the system is an ongoing research task in machine translation. Some challenges in performing domain adaptation are to decide which part of the system requires adaptation and to choose what method needs to be applied. In this paper, we used language model interpolation as a domain adaptation method and proved that it is a fast state of the art method that can be used in building adapted translation systems even when sparse domain specific material is available (i.e. especially in the case of low-resourced language pairs). The best improvement was of 15 BLEU points over the baseline system.

1 Introduction

As a response to the increased need of managing on-line available data, traditional content management system extended their functionality by offering a web-front end facility and, more recently by including cloud services. In this article we will refer to this type of system as Web Content Management System (WCMS).
Existent WCMSs focus on storage of documents in databases and provide mostly full-text search functionality. These types of systems have limited applicability, due to two reasons:

- data available online is often multilingual and
- documents within a CMS are semantically related (share some common knowledge, or belong to similar topics).

In short, currently available CMS do not exploit in production environments modern techniques from information technology like text mining, semantic Web or machine translation. Current initiatives, like the "Multilingual Web-LT" (http://www.multilingualweb.eu/), are developing now standards and best practices for dealing with multilingual content on the Web, but this is for the moment still not applied consequently for CMS.

The ICT PSP EU project ATLAS – Applied Technology for Language-Aided CMS (http://www.atlasproject.eu) – aims to fill this gap by providing three innovative Web services within a WCMS. These three Web services: i-Librarian, EUDocLib and i-Publisher are not only thematically different, but offer also different levels of intelligent information processing.

Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe, ed. by Georg Rehm, Felix Sasaki, Daniel Stein, and Andreas Witt. Volume 3, Number 1. June 2013.
ISSN 2193-6986

The ATLAS WCMS makes use of state-of-the art text-technological methods in order to extract information and cluster documents according to a given hierarchy. A text summarization module and a machine translation engine as well as a cross-lingual semantic search engine are embedded.
Currently the system is addressing six languages (Bulgarian, English, German, Greek, Polish and Romanian) from four language families. However, the chosen framework allows additions of other languages at a later point.

The focus of this paper is on the machine translation engine within the ATLAS project and on performing domain adaptation that gives significant improvements over the baseline system at evaluation. It should also be stated that the aim of the ATLAS project is to adapt state-of-the art methods in language technology with the purpose of being integrated into a content management system, thus the project is not only a research project, but also a product-oriented one. Our attention was focused on selecting the most adequate state-of-the–art method in domain adaptation for machine translation.

In natural language processing, the notion of "domain" could refer to the genre, the text type or the style of a document (Lee 2001). In this paper, we use the definition from (Plank 2011) where a domain is defined by a corpus. The problem of domain adaptation could be formulated as follows: given a large amount of bilingual source data (training data) and a small amount of target data, the purpose of the domain adaptation task is to build a system that has a good performance when evaluated on test sets that belong to the target domain. We use the terms source domain and out-of-domain interchangeably. Also, the terms target domain and in-domain are used interchangeably.

The remainder of this paper is organized as follows. In chapter two, the ATLAS Content Management System is described with details on the integration of machine translation into the ATLAS system. Chapter three presents state of the art in domain adaptation for statistical machine translation (SMT) with insight on the limitations of the current methods. The next chapter introduces the baseline translation system we used and the resources needed in order to build it. The experiments we performed in domain adaptation are presented in chapter five. We conducted two types of experiments: firstly, we identified a state of the art domain adaptation method that is easy to use and gives significant improvements over the baseline. Then, after deciding on the method, we performed various experiments on different domains from the ATLAS project and on different language pairs. The results are also presented in this chapter. The conclusions are presented in the last chapter.

2 The ATLAS Content Management System

The core online service of the ATLAS platform is i-Publisher, a powerful web-based instrument for creating, running and managing content-driven Web sites. It integrates language-based technologies to improve content navigation e.g. interlinking documents based on extracted phrases, words and names, providing short summaries and suggesting categorization concepts. Currently two different thematic content-driven Web sites, i-Librarian and EUDocLib, are being built on top of the ATLAS platform, using i-Publisher as content management layer. i-Librarian is intended to be a user-oriented Web site which allows visitors to maintain a personal workspace for storing, sharing and publishing various types of documents and to have them automatically categorized into appropriate subject categories, summarized and annotated with

important words, phrases and names. EUDocLib is planned as a publicly accessible repository of EU legal documents from the EUR-LEX collection with enhanced navigation and multilingual access.
The i-Publisher service:

- is mainly targeted at small enterprises and non-profit organizations,
- gives the ability to build content-driven Web sites via point-and-click user interface, which provide a wide set of pre-defined functionalities and whose textual content is automatically processed, i.e. categorized, summarized, annotated, etc.,
- enables publishers, information designers and graphic designers to easily collaborate,
- aims at saving authors, editors and other contributors valuable time by automatically processing textual data and allows them to work together to produce high quality content. The last evaluation round of the service indicates that users indeed see the benefit of LT-Technologies embedded into the system.

The i-Librarian service:

- addresses the needs of authors, students, young researchers and readers,
- gives the ability to easily create, organize and publish various types of documents,
- allows users to find similar documents in different languages, to share personal works with other people, and to locate the most relevant texts from large collections of unfamiliar documents.

The EUDocLib service is a particular refinement of i-Librarian targeted at the management of documents from the European Commission.

The services described above are supported through intelligent language technology components like automatic classification, named entity recognition and information extraction, automatic text summarization, machine translation and cross-lingual retrieval. These components are integrated into the system in a brick-like architecture, which means that each component is building on top of the other. The baseline brick is the language processing chains component which ensures a heterogonous linguistic processing of all documents independent of their language (Belogay, et al. 2011). A processing chain for a given language includes a number of existing tools, adjusted and/or fine-tuned to ensure their interoperability. In most respects, a language processing chain does not require development of new software modules, but rather a combination of existing tools. The basic ATLAS software[1] is distributed as a software package under GPL license. LT-plug-ins like the language processing chains or the MT-engine follows a commercial licensing. The iLibrarian is available as a Web-service and it has unrestricted access.

Machine Translation in the ATLAS System

Machine Translation is a key component of the ATLAS system. The development of the engine is particularly challenging as the translation should be used in different domains. Additionally, the considered language-pairs belong to the low resourced group[2], for which bilingual training and test material is available in limited amount.

[1] http://atlasproject.eu
[2] See http://www.meta-net.eu/whitepapers.

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

- for i-Publisher Service (generic platform for generating websites) 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 human post processing;
- for i-Librarian and EuDocLib (dedicated Web services for collecting documents) 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. If the translation is considered acceptable it will be stored into a database.

The integration of a machine translation engine into a web-based content management system in general, and into the ATLAS system in particular, presents several challenges from the user point of view among which we mention two challenges that were dealt within the ATLAS System:

1. The user may retrieve documents from different domains. Domain adaptation 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 modern languages do not work equally well 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.

For the MT-Engine of the ATLAS system we decided on a hybrid architecture combining EBMT (Gavrila 2011) and SMT (Koehn, Hoang, et al. 2007) at phrase-based level (no syntactic trees will be used). 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 a way that ellipses and anaphora are omitted, and lexical material is adapted to the training corpus.

The information extraction module is providing information about metadata of the document including publication age. For documents previous to 1900 we will not provide translation, explaining to 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. The described architecture is presented in Figure 1.

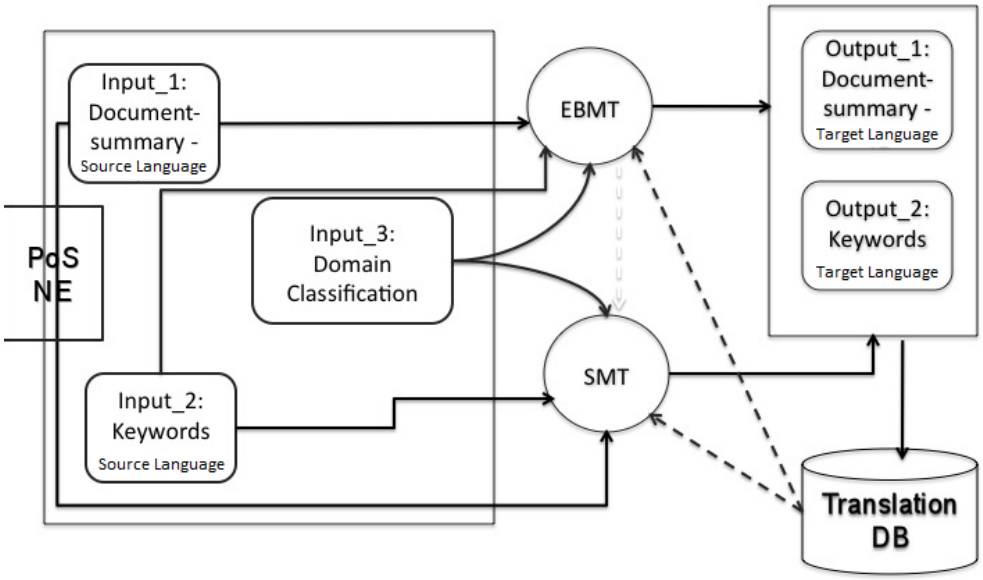

Figure 1: System architecture for the ATLAS-Engine

In order to perform domain adaptation we collected domain specific corpora for 13 upper domains in the categorization tree embedded in the ATLAS system and performed various experiments to choose a fast and easy to use domain adaptation method that can significantly improve the translation.

3 State of the art in domain adaptation for Statistical Machine Translation

Domain adaptation (DA) can be classified by taking into consideration the models that are adapted, the resources that are used or the type of supervision used.

In the following table, multiple types of approaches are presented. The numbers of the papers that appear after the table are given in the column *Reference* according to the approach the paper uses in adaptation.

Classification of Domain Adaptation approaches for SMT		
Approach	**Type**	**Reference**
Model	▪ Word alignment model ▪ Language model ▪ Translation model ▪ Reordering model	2 1, 3, 4, 6, 8 3, 4, 5, 7, 8 4, 7, 9
Resources	▪ Monolingual corpora ▪ Parallel corpora ▪ Comparable corpora ▪ Web-crawled data	5, 6, 7 1, 2, 3, 4, 6, 9 5 8
Supervision	▪ Supervised ▪ Unsupervised ▪ Semi-supervised	1, 2, 3, 4, 8, 9 5, 7 6

Table 1: Domain adaptation approaches

In the following, the state of the art in domain adaptation for statistical machine translation (SMT) is presented using the chronological order by year of papers published. All papers evaluated their methods using one or more evaluation metrics and the most common metric used was BLEU (Papineni, Roukos, et al. 2002).

1. An *unsupervised language model adaptation* method is explored in (Zhao, Eck and Vogel 2004) where structured query models are used. Translations are obtained using a baseline translation system that uses a general language model. Then the hypotheses from the output are converted into queries with the aim of retrieving similar sentences from very large news documents collections. Using these retrieved sentences a language model is built and linearly interpolated with the baseline language model. The final step consists in using the interpolated language model to produce new translations.
2. Experiments in *alignment adaptation* were described in (Wu, Wang and Liu 2005) where out-of-domain data is used in order to get better results when performing 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 (size of out-of-domain >> size of in-domain). A new alignment model results by interpolating the two models.
3. Multiple experiments in domain adaptation for SMT were explored by (Koehn and Schroeder 2007). The baseline systems were trained using different methods: using only out-of-domain data, using only in-domain data and using concatenated out-of-domain and in-domain data. Among these three baselines the best BLEU score was obtained using the concatenated data. The adaptation methods used are: use only the in-domain data to build the language model, *interpolate the LM* estimated from out-of-domain data with the LM estimated from in-domain data, *use both language models as separate features* with weights set using MERT and the last method makes use of *factored translation models* where two decoding paths corresponding to each translation table are used.
4. In (Chen, Zhang, Aw and Li 2008) n-best hypotheses are used fo*r 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.
5. One approach to *translation model adaption* relies on using *comparable corpora*[3]. In (Snover, Dorr and 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 a similar topic as the source language document that will be translated. The documents are used to increase the probability of generating texts that are similar to the comparable document.
6. The use of a *domain dictionary* and *monolingual corpora* is employed in (Wu, Wang and Zong 2008). The out-of-domain data is used in estimating a language model and constructing a phrase table, probabilities are assigned to entries in the in-domain translation dictionary, a phrase table for the in-domain is constructed, and the two phrase tables are combined. If in-domain target data is available, a language model is estimated and *combined* with the out-of-domain one. If in-domain source data is

[3] Texts that have the same topic and similar content.

available, the already built model is used in translating the data thus obtaining a synthetic corpus that is added to the training data.

7. *Monolingual resources* are also explored in (Bertoldi and Federico 2009). The approaches pursued are: use baseline translation system to generate synthetic bilingual data, use the generated data for *translation and reordering model adaptation* and use synthetic text or given target texts for language model adaptation.
8. Recent work in Domain Adaptation for Statistical Machine Translation focus on using *web-crawled data for building language models, improving translation models, tuning and testing*. In (Pecina, Toral and Way, et al. 2011) and (Pecina, Toral and Papavassiliou, et al. 2012), domain-specific data is obtained by web-crawling. The basic workflow of their work is: use focused web-crawling, text normalization, language identification, document clean-up and near-duplicate detection.
9. (Ling, Luis, Graca, Coheur and 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.

In the following figure, a domain adaptation setup for statistical machine translation is presented.

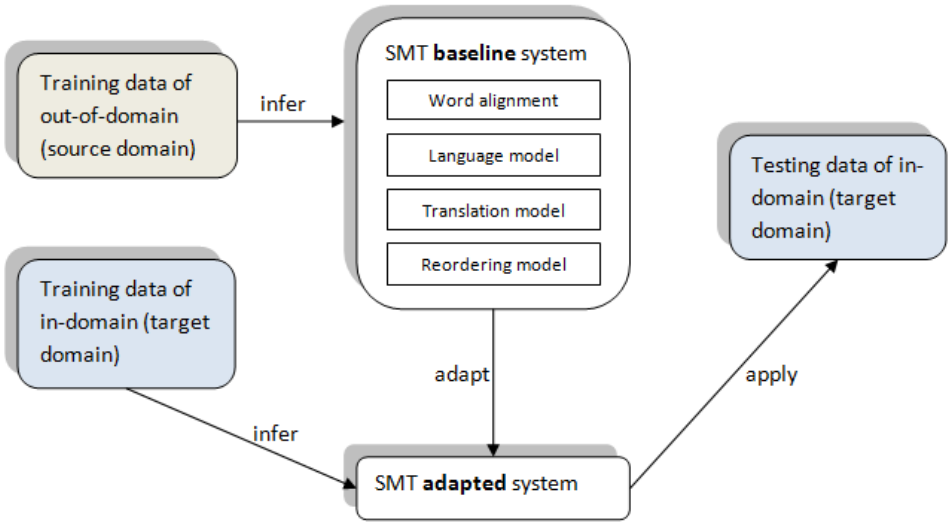

Figure 2: Domain adaptation setup[4]

4 The baseline translation system

The experiments were run using the widely-used open-source toolkit Moses[5]. Moses is a statistical machine translation system, which utilizes large parallel corpora in order to train the translation system. We used in our experiments the phrase-based translation

[4] Figure adapted from (Plank 2011) where a DA setup is presented in the task of parser adaptation. The adapted system is made up of the same type of models as the baseline system, but these models were ommited in the drawing due to the fact that one or more models can be adapted.

[5] http://www.statmt.org/moses/index.php?n=Main.HomePage

model made accessible by the Moses system. The training pipeline[6] consists of the following steps: pre-processing data by tokenizing, true casing and cleaning using tools from the Moses toolkit, followed by language model training and translation training where a word-alignment is performed, phrases are extracted and multiple scores are computed. For the language model training, we chose the SRILM[7] toolkit, which is also open-source. It builds statistical language models and it also offers the possibility of interpolating language models. As for the word-alignments, they were performed using GIZA++[8], a commonly used tool for word alignments. Because of the fact that this tool runs slowly on long sentences or fails in aligning them, we chose to work with a maximum sentence length of 50 words.

In order to train a statistical machine translation system, parallel corpora were needed. The corpus JRC-Acquis[9] is a multilingual parallel corpus for 22 European languages consisting of paragraph alignments for 231 pairs[10] of languages. The data is made up of a selection of European documents referred to as Acquis. This term identifies the body of common rights and obligations that bind all the member states from the European Union. The choice of using this corpus is motivated by the fact that it is freely available, it has a large dimension and it contains aligned corpora for all the language pairs within the ATLAS project.

The experiments were evaluated using the common evaluation metric BLEU which uses counts of n-grams.

5 Experiments in Domain Adaptation

In order to investigate current methods of domain adaptation, experiments were performed that were inspired by the work presented in (Koehn and Schroeder 2007). In their work, the language pair French-English was used, while the Europarl corpus was used as out-of-domain. The in-domain was made up of the News Commentary corpus. The BLEU scores for each of the adaptation methods proposed are presented below.

Method	BLEU
Large out-of-domain training data	25.11
Small in-domain training data	25.88
Combined training data	26.69
In-domain language model	27.46
Language model interpolation	27.12
Two language models	27.30
Two translation models	27.64

Table 2: BLEU scores for the experiments from (Koehn and Schroeder 2007)

From the seven experiments conducted by (Koehn and Schroeder 2007), we selected three experiments that can be easily reproduced (combined training data, in-domain language model and interpolated language model). Then we identified the best one according to the BLEU scores, which was the in-domain language model method.

[6] http://www.statmt.org/moses/?n=Moses.Baseline
[7] http://www.speech.sri.com/projects/srilm/download.html
[8] http://code.google.com/p/giza-pp/
[9] http://ipsc.jrc.ec.europa.eu/index.php?id=198
[10] http://langtech.jrc.ec.europa.eu/Documents/070622_Poster_JRC-Acquis.pdf

We performed three experiments using the out-of-domain JRC-Acquis, the in-domain Politics from the parallel corpora ATLAS and the language pair Bulgarian-English. Even though the out-of-domain and the in-domain both belong to the same topic, they differ in the text style. The aim of these experiments was to verify if using the in-domain language model method is also the best adaptation method for our settings. But, as results show in Table 4, the best method is actually language model interpolation (even though using only the in-domain language model gives close results to the language model interpolation).

In the following tables, the statistics for the corpora used and the BLEU results are presented.

# sentences in-domain Politics	# sentences out-of-domain JRC-Aquis	# sentences test-set (Politics domain)
56796	306767	3000

Table 3: Statistics for the corpora used in the experiments for BG-EN

Method	BLEU
Combined training data	24.98
In-domain language model	39.07
Language model interpolation	39.36

Table 4: BLEU results for the adaptation methods tested on BG-EN with the in-domain Politics

In order to estimate language models and to perform language model (LM) interpolation, we used the SRILM toolkit. Two language models were built: one for the target language estimated from the out-of-domain corpus and one for the target language estimated from the in-domain corpus. Then, we used the *compute-best-mix* script from SRILM to compute the best interpolation weight. This weight and the two language models were used in order to build the interpolated language model.

Lang. Pair	BLEU Adapted System (AS)	BLEU Baseline System (BS)	# sent. In-domain Corpus	# sent. Out-of-domain Corpus	# sent. Test Set	improvement
DE-EN	13.18	9.84	93160	1199447	4500	3.34
EN-DE	11.3	7.96	93160	1199447	4500	3.34
EN-RO	14.97	6.98	10109	336455	500	7.99
RO-BG	19.58	7.22	10410	241670	500	12.36
RO-EN	23.82	9.69	10109	336455	500	14.13

Table 5: Results of experiments on Business in-domain

After deciding what was the best adaptation method in our current settings (LM interpolation), we conducted experiments on other ATLAS in-domain corpora: Sociology and Business. We wanted to check the correlation between the size of the out-of-domain, the in-domain and the improvement[11] on different language pairs: English-German, German-English, Romanian-English, English-Romanian and Romanian-Bulgarian. As can be seen in Table 5 and Table 6, there is a big difference between the

[11] We use the term "improvement" to define the difference between the BLEU score of the adapted system and the BLEU score of the baseline system.

sizes of the Business in-domain and the Sociology in-domain. Another goal of our work was to evaluate the chosen DA method, by comparing the BLEU scores of the baseline systems to the scores of the adapted systems.

The test sets belonged to the same domain as the in-domain corpus and the size of the test sets was set to approximately 5% of the size of the in-domain corpora.

Lang. Pair	BLEU Adapted System (AS)	BLEU Baseline System (BS)	#sent. In-domain Corpus	#sent. Out-of-domain Corpus	#sent. Test Set	improvement
EN-DE	30.05	22.3	1808	1199447	100	7.75
DE-EN	35.21	27.3	1808	1199447	100	7.91
EN-RO	30.46	21.92	2010	336455	100	8.54
RO-BG	17.68	7.31	2176	241670	100	10.37
RO-EN	36.82	21.71	2010	336455	100	15.11

Table 6: Results of experiments on Sociology in-domain

We observed from our experiments that there is a correlation between the size of the in-domain corpus, the out-of-domain corpus, the number of test sentences and the BLEU score. On the Sociology experiments, the size of test sets is set to 100 sentences and the size of the in-domain data is between 1800 and 2200 sentences. Even though the size of the in-domain data for RO-BG is similar to the size of the in-domain data for RO-EN, the size of the out-of-domains for the two language pairs differs by almost 100000 sentences. This is the reason why there is a large difference in BLEU scores for the two systems (10.37 for RO-BG and 15.11 for RO-EN). The same correlations can be observed on the Business domain (12.36 for RO-BG and 14.13 for RO-EN).

While the most significant improvement among all ten experiments was on the in-domain Sociology, language pair RO-EN (BLEU difference of 15.11), the less significant improvement of 3.34 BLEU points was made on the Business domain for the language pairs EN-DE and DE-EN. The reason for this small improvement lies in the large amounts of data used for the in-domain and also for the out-of-domain corpora. Sentence alignment problems appear in large corpora leading to word-alignment problems and in the end, problems in the translation, which result in low BLEU scores.

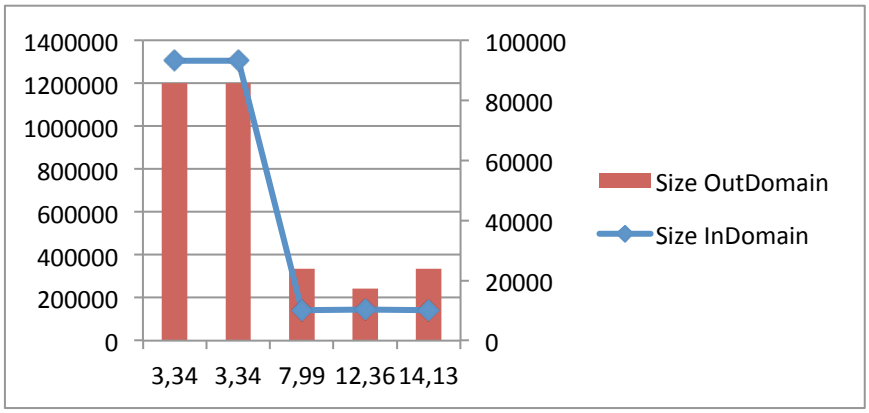

Figure 3: Improvement for the experiments using in-domain Business

In Figure 3 we plotted on the X axis the improvement, on the left Y axis the size of the out-of-domain and on the second Y axis, the size of the in-domain. It can be

observed that for the experiments that used large amounts of both out-of-domain and in-domain data, the improvement was the lowest. When the out-of-domain corpus and the in-domain corpus had smaller dimensions, the improvement was significantly better. Another case, large out-domain corpus and small in-domain corpus, can be observed in Figure 4, where all ten experiments are illustrated. In this case, the improvement is also significant.

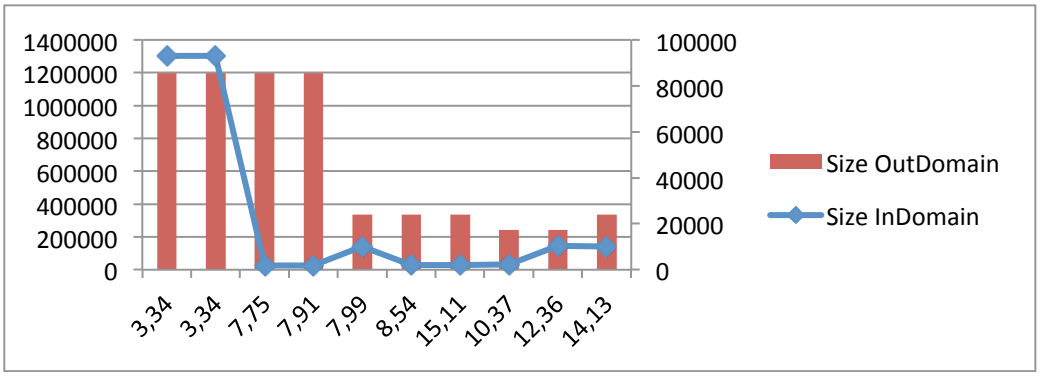

Figure 4: Improvement for all experiments

By looking at the improvements, we came to the conclusion that having more in-domain data does not necessarily lead to better results and that the chosen adaptation method is more important than the amount of in-domain data.

In the following table, an example of translations from the in-domain Sociology and the language pair Romanian-English is presented. This is the experiment that gave the best improvement among all experiments (15.11). In the sentence translated using the baseline system, the unknown words are underlined. The adapted system could translate all the words in this case and the sense of the sentence is similar to the sense of the reference sentence.

Type	**Sentence**
Source	toate declarațiile de susținere vor fi distruse în termen de 18 luni de la data de înregistrare a inițiativei propuse de cetățeni , sau , în cazul unor proceduri administrative sau juridice , cel târziu la o săptămână după data încheierii procedurilor în cauză .
Reference	all statements of support will be destroyed at the latest 18 months after the date of registration of the proposed citizens ' initiative , or , in the case of administrative or legal proceedings , at the latest one week after the date of conclusion of the said proceedings .
Adapted System	all statements of support will be destroyed 18 months after the registration of initiative proposed by citizens , or , in the case of administrative procedures or legal , at the latest one week after the date of the procedures in question .
Baseline System	all declarațiile of susținere shall be destroyed within 18 months from the date of registration of inițiativei proposed by cetățeni , or , in the case of administrative or legal , not later than one week from the date of conclusion of the procedures in question .

Table 7: Translation example using a test set sentence that belongs to the Sociology domain, RO-EN

The next example is taken from a test set belonging to the Business domain, language pair German-English. This is the experiment that gave the lowest improvement among all experiments (3.34). Even though in the sentence translated by the adapted system there are no unknown words, the sense of the sentence is not very close to the sense of the reference sentence.

Type	Sentence
Source	eine solche anbindung birgt das risiko , dass aufwärtsgerichtete inflationsschocks zu einer lohn-preis-spirale führen , was sich in den betroffenen ländern nachteilig auf beschäftigung und wettbewerbsfähigkeit auswirken würde .
Reference	such schemes involve the risk of upward shocks in inflation leading to a wage-price spiral , which would be detrimental to employment and competitiveness in the countries concerned .
Adapted System	such carries the risk that monetary policy discussion of an early , in the countries concerned detrimental to employment and competitiveness .
Baseline System	such a link between carries the risk that aufwärtsgerichtete inflationsschocks lead to a lohn-preis-spirale , in the countries concerned on employment and competitiveness .

Table 8: Translation example using a test set sentence that belongs to the Business domain, DE-EN

6 Conclusions

In this paper we presented the ATLAS Content Management System focusing on the integration of machine translation into the system. A current problem of machine translation is domain adaptation as many statistical systems are trained on a general domain and used on divergent domains. We have investigated three methods presented in (Koehn and Schroeder 2007) in order to choose a domain adaptation method that can be easily and fast integrated into the system. The best adaptation method[12] among these three was the usage of in-domain language model. However, our experiments show that in our current settings, the best method is language model adaptation.

Afterwards, we wanted to evaluate the chosen DA method. For this reason we performed experiments using baseline systems trained on JRC-Acquis and evaluated them using BLEU. In order to perform domain adaptation, we used the Business and the Sociology in-domains and the language pairs German-English, English-German, Romanian-Bulgarian, English-Romanian and Romanian-English. The BLEU scores for all the adapted systems outperformed the BLEU scores of the baseline systems. It is important to emphasize the high BLEU differences between the baseline systems and the adapted systems (the best improvement was of 15.11 BLEU points).

Two important ideas are highlighted by the results of our experiments. When performing domain adaptation it is not necessary to have a large in-domain corpus in order to attain good adaptation results (a size of 2000 sentences is sufficient). The other conclusion is that in our current settings, choosing the method of adaptation is more important that having a large in-domain corpus.

We conclude that having in-domain data is important for domain adaptation, but it is more important to choose a good adaptation method that gives significant improvements when applied to different in-domains and different language pairs.

7 Acknowledgements

ATLAS is a project funded by the European Commission under the CIP ICT Policy Support Program.

We want to thank the anonymous reviewers for their comments and constructive suggestions.

[12] In our settings

8 References

Bellegarda, Jerome. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*. Vol. 42, 93-108.

Belogay, Anelia, Dan Cristea, Eugen Ignat, Diman Karagiozov, Koeva Svetla, Maciej Ogrodniczuk, Adam Przepiórkowski, Przepiórkowski Raxis, and Cristina Vertan. 2011. Language processing chains in ATLAS. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics.*

Bertoldi, Nicola, and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. *Proceedings of the 4th Workshop on Statistical Machine Translation.* 182-189.

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 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.* 17-24.

Chen, Boxing, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting N-best Hypotheses for SMT Self-Enhancement. *Proceedings of ACL-08: HLT, Short Papers (Companion Volume).* 157-160.

Gavrila, Monica. 2011. Constrained Recombination in an Example-based Machine Translation System. *Proceedings of the EAMT-2011: the 15th Annual Conference of the European Association for Machine Translation.*

Ker, Sue J., and Jason S. Chang. 1997. A class-based approach to word alignment. *Computational Linguistics*. Vol. 23, 313-343.

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

Koehn, Philipp, and Josh Schroeder. 2007. Experiments in Domain Adaptation for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation.* 224-227.

Lee, David YW. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language, Learning and Technoloy*. Vol. 5, No. 3, 37-72.

Ling, Wang, Tiago Luis, Joao Graca, Luisa Coheur, and Isabel Trancoso. 2011. Reordering Modeling using Weighted Alignment Matrices. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.* Vol. 2, 450-454.

Och, Franz Josef, and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*. 19-51.

Och, Franz Josef, and Hermann Ney. 2000. Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics.* 440-447.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* 311-318.

Pecina, Pavel, Antonio Toral, Andy Way, Vassilis Papavassiliou, Prokopis Prokopidis, and Maria Giagkou. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. *Proceedings of the 15th Conference of the European Association for Machine Translation.*

Pecina, Pavel, Antonio Toral, Vassilis Papavassiliou, Prokopis Prokopidis, and Josef van Genabith. 2012. Domain Adaptation of Statistical Machine Translation using Web-crawled resources: A Case Study. *Proceedings of the 16th EAMT Conference.*

Plank, Barbara. 2011. Dissertation "Domain Adaptation for Parsing" - Chapter 3: Domain Adaptation. *Groningen Dissertations in Linguistics 96*. ISSN 0928-0030. http://dissertations.ub.rug.nl/FILES/faculties/arts/2011/b.plank/03c3.pdf (retrieved April 2013).

Snover, Matthew, Bonnie Dorr, and Richard Schwartz. 2008. Language and Translation Model Adaptation using Comparable Corpora. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.*

Stolcke, Andreas. 2002. SRILM - An Extensible Language Modeling Toolkit. *Proceedings of the International Conference on Spoken Language Processing.* Vol. 2, 901-904.

Wu, Hua, Haifeng Wang, and Chengquing Zong. 2008. Domain Adaptation for Statistical Machine Translation with Domain Dictionary and Monolingual Resources. *Proceedings of the 22nd International Conference on Computational Linguistics.* 993-1000.

Wu, Hua, Haifeng Wang, and Zhanyi Liu. 2005. Alignment Model Adaptation for Domain-Specific Word Alignment. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics.* 467-474.

Zhao, Bing, Matthias Eck, and Stephan Vogel. 2004. Language Model Adaptation for Statistical Machine Translation with Structured Query Models. *Proceedings of Coling 2004.* 411-417.