Using Manual and Parallel Aligned Corpora for Machine Translation Services within an On-line Content Management System

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

Monica Gavrila
University of Hamburg
gavrila@informatik.uni-hamburg.de

Abstract

Web content management systems (WCMSs) are a popular instrument for gathering, navigating and assessing information in environments such as Digital Libraries or e-Learning. Such environments are characterized not only through a critical amount of documents, but also by their domain heterogeneity, relative to format, domain or date of production, and their multilingual character. Methods from Information and Language Technology are the "plug-ins" necessary to any WCMS in order to ensure a proper functionality, given the features mentioned above. Among these "plug-ins", machine translation (MT) is a key component, which enables translation of meta-data and content either for the user or for other components of the WCMS (i.e. cross-lingual retrieval component). However, the MT task is extremely challenging and lacks frequently the availability of adequate training data. In this paper we will present a WCMS including machine translation, explain the related MT challenges, and discuss the employment of corpora as training material, which are manually and automatically parallel aligned.

1 Introduction

During the last couple of years, the number of applications which are entirely Web-based or offer at least some Web front-ends has grown dramatically. As a response to the need of managing all this data, a new type of systems appeared: the web-content management systems. In this article we will refer to this type of systems as WCMS. Existent WCMSs focus on storage of documents in databases and provide mostly full-text search functionalities. These types of systems have limited applicability, due to reasons such as the following:

- data available on-line is often multilingual;
- documents within a content management system (CMS) are semantically related (share some common knowledge or belong to similar topics).

Shortly, currently available CMSs do not exploit modern techniques from information technology like text mining, semantic web or machine translation.

The recently launched ICT PSP EU project ATLAS (Applied Technology for Language-Aided CMS[1]) aims to fill in 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 also offer different levels of intelligent information processing.

The ATLAS WCMS makes use of state-of-the-art text technology 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. The system is addressing for the moment seven languages (Bulgarian, Croatian, English, German, Greek, Polish and Romanian) from four different language families. However, the chosen framework allows additions of new languages at a later point.

Machine Translation is a key component of the ATLAS-WCMS and it will be embedded in all three services of the system. The development of the engine is particularly challenging as the translation should be used in different domains and on

[1] http://www.atlasproject.eu.

Proceedings of Recent Advances in Natural Language Processing, pages 53–58,
Hissar, Bulgaria, 12-14 September 2011.

different text-genres. Additionally, the considered language-pairs belong most of them to the lesser resourced group of languages, for which bilingual training and test material is available only in limited amount.

The availability of adequate and comparable training data for all language pairs in the ATLAS system played an important role in the architectural design of the MT-engine. The selection of training data was preceded by experiments on selected language pairs. Through these experiments we intended to investigate if small parallel corpora can be also used and with which implications on the translation quality. We investigated additionally the automatic (sentence) alignment in larger corpora in order to understand which implications alignment errors may have on the translation process.

In the following sections we report about our findings as follows: in Section 2 we present briefly the ATLAS functionality and describe the corresponding challenges for the machine translation engine. In section 3 we present the data we used for experiments and analyze it from the linguistic point of view. Section 4 deals with experiments which investigate the dependency between the amount of the training data and the translation quality. Section 5 gives an overview of future experiments and implementation steps.

2 MT-challenges in the ATLAS-System

2.1 The ATLAS-System

The core on-line 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. by interlinking documents based on extracted phrases, words and names, providing short summaries and suggested categorization concepts. Currently two different thematic content-driven Web sites are being built on top of ATLAS platform, using i-Publisher as content management layer: i-Librarian and EUDocLib. 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 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. All three services operate in the multilingual setting described in Section 1. To justify the need of embedded language technology tools within the ATLAS platform we detail here only the functionalities of i-Librarian.

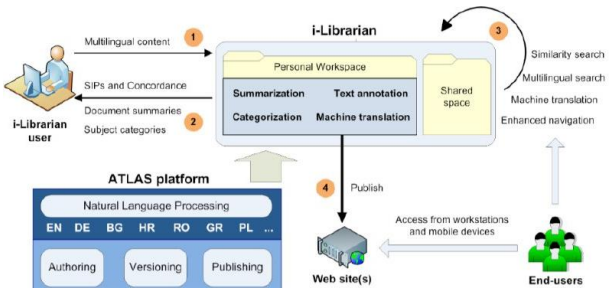

Figure 1: The iLibrarian Architecture

The i-Librarian service (see Figure 2.1):

- 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 essential texts from large collections of unfamiliar documents.

The facilities 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 built on top of the other. The baseline brick is the language processing chains component which ensure a heterogeneous linguistic processing of all documents independent of their language (Ogrodniczuk, 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 combining existing tools.

With respect to the machine translation engine the language processing tools provide the

part-of-speech (PoS) annotation necessary for factored models and ensure named entity recognition. Other bricks of the ATLAS architecture feed information into the translation engine as follows:

1. the document categorization gives information about the domain of a particular document;

2. the automatic summarization deals with anaphora resolutions and pre-processes the document in order to simplify the translation task.

2.2 Challenges of the MT-Task

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

- the MT-engine is serving as a translation aid tool for publishing multilingual content for i-Publisher. Text is submitted to the translation engine and the result is subject to the human post processing;

- for i-Librarian and EuDocLib, 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 wants to store them. If the translation is considered as acceptable, it will be stored into a database.

The integration of a machine translation engine into a web based content management system, presents from the user point of view two main challenges:

- 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;

- 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

	BG	DE	EN	GR	HR	PL	RO
BG		666 k	161 k+ 647 k	169 k+ 749 k	165 k	663 k	168 k+ 361 k
DE	ACQUIS		1581 k + 9 k + 1264 k+ 42 k	TBC + 9 k + 1132 k		9 k+ 1271 k+ 37 k	392 k+ 33 k
EN	SETIMES ACQUIS	EUROPARL EUConst ACQUIS PHP		960 k + 10 k+ 1083 k	150 k	10 k+ 1259 k+ 40 k	173 k+ 391 k+ 36 k
GR	SETIMES ACQUIS	EUROPARL Euconst ACQUIS	EUROPARL SETIMES EUConst ACQUIS		160 k	10 k + 1115 k	176 k+ 337 k
HR	SETIMES		SETIMES	SETIMES			176 k
PL	ACQUIS	EUConst ACQUIS PHP	EUConst ACQUIS PHP	EUConst ACQUIS			398 k+ 42 k
RO	SETIMES ACQUIS	ACQUIS PHP	SETIMES ACQUIS PHP	SETIMES ACQUIS	SETIMES	ACQUIS PHP	

Figure 2: Available parallel corpora for all language pairs within the ATLAS system.

domain and language variation independent and works for a couple of heterogeneous language pairs. Therefore our approach envisage a system of user guidance, so that the availability and the foreseen system-performance is transparent at any time.

From the development point of view the main challenge is provided by the high number of language pairs[2], most of them involving languages with rich morphology and belonging to structural different language families. For most of the language pairs a limited number of parallel aligned corpora are available. Additionally, the ATLAS platform should provide a basic comparable functionality for all language pair, so we cannot train models for different language pairs on completely different corpora.

After collecting information regarding parallel corpora for all involved language pairs, we decided to focus the development of basic training models on those summarized in Figure 2.2[3].

It can be observed that with exception of Croatian, for all other involved languages the JRC-Acquis[4] corpus offers a good training basis (coverage and size). In order to ensure domain portability we decided to train domain factored models as in (Niehues and Waibel, 2010). This approach allows the usage of small domain specific corpora. Small corpora have the advantage that they can be manually aligned, or at least manually corrected. In order to see how the translation engine behaves when exposed to large but automatically trained corpora and to small but manually aligned texts, we performed several analyses described in sec-

[2]More than 40 language-pairs.

[3]We do not consider in this table the recent additions from February 4th, 2011 concerning the Europarl corpus.

[4]`http://optima.jrc.it/Acquis/`.

tion 4.

3 Manually Aligned Small Corpora vs. Automatically Aligned Large Corpora

We decided to make selective experiments on corpora involving following language pairs: English, Romanian and German. Our choice is based on the availability of human evaluators speaking all three languages, but also by the fact that the languages belong to structural different families (Romania is in the Latin language family, English and German are Germanic languages). Additionally Romanian and German are highly inflected.

3.1 JRC-Acquis

The JRC-Acquis Communautaire is nowadays one of the mostly used parallel aligned corpus for training models in statistical machine translation (Koehn et al., 2009). We do not make here an extensive presentation of the SMT system but present in Table 1 and 2 just a comparative statistics on the three selected languages[5]. From these tables we can infer that the size of the training material has large variations across different language pairs within the JRC-Acquis.

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: JRC-Acquis alignment statistics (docs=documents).

The corpus is automatically paragraph-aligned, where a paragraph is a simple or complex sentence or a sub-sentential phrase (such as noun-phrase).

3.2 RoGER

RoGER (Romanian German English, Russian) is a parallel corpus, manually 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 automatically transformed into text files (.RTF), where 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. This way we obtained two alignment files. Finally, the two alignment files obtained were merged, so that, after all, RoGER contained all four languages. The merged text files are XML encoded, as shown below:

```
  <?xml version=''1.0''
encoding=''UTF-8''?>
<sentences>
.................
<sentence id=''1010''>
<en>Press Options and some of the
following options may be available
.</en>
<de>Druecken Sie Optionen .  und
einige der folgenden Optionen sind ggf.
verfuegbar .</de>
<ro>Apasati Optiuni dupa care unele din
urmatoarele optiuni pot fi disponibile
.</ro>
<ru>...</ru>
</sentence>
.......................
</sentences>
```

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. Punctuation signs are considered tokens. More about the RoGER corpus can found in (Gavrila and Elita, 2006)

3.3 Linguistic Analysis of the Corpora

From both corpora we randomly extracted about 100 sentences, i.e. 100 sentences from the JRC-Acquis corpus for Romanian-English and 100 sentences from the RoGER corpus and the same language pair and direction of translation. These sentences were analyzed with respect to translation divergences and translation mismatches.

Translation divergence means that the same information appears in both SL and TL, but the structure of the sentence is different. Translation

[5]Information source: `http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/`.

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

Table 1: JRC-Acquis statistics.

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: Statistics on RoGER.

divergences are presented in the literature in (Dorr et al., 1999) and (Dorr, 1994). In the case of a translation mismatch the information that can be extracted from the SL and TL sentence is not the same. Translation mismatches have received less attention in the literature (Kameyama et al., 1991), but for corpus-based approaches they are important, as they directly influence the translation process.

Following translation challenges were observed within the JRC-Acquis:

- Divergences
  - Noun (NN) - adjective (Adj) inversion
  - Noun-Preposition-Noun (NN-prep-NN) translated as adjective-Noun (Adj- NN)
  - Subordinate clause translated as adjective
  - Different argument structure
  - Different type of articles
  - Voice change (for verbs)
- Mismatches
  - Extra information (the TL sentence is more explicit than the SL one)
  - Reformulations
- Wrong translation (due to incorrect alignment)

All these phenomena have a direct (negative) influence on the automatic evaluation scores. Although the corpus is domain restricted, the likelihood of at least one divergence or mismatch type occurring in a sentence is high. Only in approximately 10% of the sentences no phenomenon was encountered. As we encountered totally wrong translations in the corpus, it shows that the (paragraph-) alignments in JRC-Acquis are not always correct.

We also analyzed 100 sentences from the center of the RoGER corpus. We noticed that the diversity of the challenges is reduced, while the number of challenges is sometimes higher compared to what had been encountered in JRC-Acquis, with up to five challenges in an example (a sentence and its translation). Usually there is a one-to-one translation. Only in 12% of cases additional information appeared for one of the languages and in only 9% reformulations have been used. Two phenomena have been found most often: NN-prep-NN translated as NN-NN (or Adj-NN) and Adj-NN inversions.

3.4 JRC-Acquis vs. RoGER

The average number of challenges in JRC-Acquis (1.89 challenges per sentences) is lower than the average number in RoGER (2.20 challenges per sentence) for the languages analyzed. However, challenges with a more negative impact on the translation quality (such as "Wrong translation"or "Reformulations") appear more frequently in JRC-Acquis. The phenomenon encountered more often for the language-pair analyzed is noun-adjective inversions.

4 Implications on the Design of the MT-Engine in ATLAS

The MT-Engine within the ATLAS System follows the hybrid approach combining a statistical based component and an example-based one. Both approaches are highly dependent from the quality

and size of the training data The linguistic analysis above shows that both corpora present translation challenges which influence negative any further automatic processing. Therefore we argue that small domain specific corpora should be aligned manually at sentence level, or at least the alignment has to be checked manually.

Additional experiments presented in (Gavrila and Vertan., 2011) shown that using ROGER as training and test corpus, the performance of the system does not decrease dramatically. Our explanation relies on the linguistic observations in Section 3. The linguistic challenges are balanced by the manual alignment. In this way the corpus, although small has a more correct sentence alignment which triggers a more correct word alignment.

These experiments lead to the conclusion that for the ATLAS-System:

- JRC-Acquis will be used as basis training corpus, without making an manual corrections. This is impossible by the size of the corpus
- Small domain specific corpora will be first manually aligned at sentence level and afterwards injected in domain factored models.

5 Conclusion and Further Work

In this paper we described the integration of a machine translation engine within a WCMS system, dealing with a large number of less resourced languages. We investigated the linguistic characteristics of two parallel corpora and show how these influence the translation quality. Further work concerns a statistical relevant analysis of the linguistic phenomena presented in Section 3, involving other manually built corpora and other language-pairs.

Acknowledgments

The present work contains ideas from the ATLAS EU-Project, supported through the ICT-PSP-Programme of the EU-Commission (Topic "Multilingual Web")(Sections 2 and 4) and from Monica Gavrila's Ph.D research conducted at the University of Hamburg (Section 3).

References

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

Bonnie J. Dorr. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633, December.

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

Monica Gavrila and Cristina Vertan. 2011. Training data in statistical machine translation - the more, the better? In *Proceedings of the RANLP-2011 Conference*, Hissar, Bulgaria, September.

Megumi Kameyama, Ryo Ochitani, and Stanley Peters. 1991. Resolving translation mismatches with information flow. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 193–200, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe. In *Proceedings of the MT Summit XII*, pages 65–72, Ottawa, Canada, August.

Jan Niehues and Alex Waibel. 2010. Domain adaptation in statistical machine translation using factored translation models. In *Proceedings of EAMT 2010*, Saint-Raphael.

Maciej Ogrodniczuk. 2011. I-publisher, i-librarian and eudoclib linguistic services for the web. In *Proceedings of the PALC 2011 Conference*.