

# Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation

Inguna Skadiņa

Tilde

Vienības gatve 75a, Rīga, LATVIA

E-mail: inguna.skadina@tilde.lv

## Abstract

This abstract presents the FP7 project ACCURAT that aims to research methods and create tools that find, measure, and use bi/multilingual comparable corpora to improve the quality of machine translation for under-resourced languages and narrow domains. Work on corpora collection, assessment of the comparability of documents pairs in collected corpora, extraction of parallel data for the machine translation (MT) task, and application to the MT task is presented.

**Keywords:** comparable corpora, under-resourced languages, comparability metric, information extraction, machine translation

## 1. Introduction

The applicability of current data-driven methods directly depends on the availability of very large quantities of parallel corpus data. For this reason, the translation quality of data-driven MT systems varies dramatically from being quite good for language pairs and domains with large corpora available to being almost unusable for under-resourced languages and domains.

The ACCURAT project (*Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation*) addresses this issue by developing technology for using comparable corpora as resources for machine translation systems (Skadiņa et al., 2010a; Eisele and Xu, 2010). The project aims to research methodology and tools that measure, find, and use comparable corpora to improve the quality of MT for under-resourced languages and narrow domains (e.g., renewable energy and topical news).

The objectives of the ACCURAT project are to:

- Research methods for automatic acquisition of a comparable corpus from the Web which can be used as a source to extract data for MT;
- Create comparability metrics – to develop the methodology and determine criteria to measure the similarity of source and target language documents in comparable corpora;
- Research methods and develop tools for the alignment and extraction of lexical, terminological, and other linguistic data from comparable corpora;
- Measure improvements achieved by applying acquired data against baseline results from statistical and rule-based machine translation systems.

The ACCURAT project particularly targets a number of under-resourced languages: Croatian, Estonian, Greek, Latvian, Lithuanian, and Romanian.

This abstract provides an overview of research results and the tools developed within the project to achieve the above mentioned objectives (more details can be found in Skadiņa et al., 2012 and papers of consortium partners listed in References).

## 2. Methods for building a comparable corpus from the Web

Several novel approaches how to build a comparable corpus from the Web that are applicable to under-resourced languages have been researched. Tools for the identification of comparable documents in Wikipedia, news documents, and narrow domains have been developed.

For **news texts**, a two-stage method that first gathers documents monolingually and then pairs them across languages to build a comparable corpus has been developed (Aker et al., 2012). In the gathering stage, news texts are downloaded separately in each project language at regular intervals from Google News. The titles are used in further queries for gathering more related articles from Google News. To overcome the relative scarcity of news in non-English languages, titles from the English news articles are parsed for named entities which are then translated into the non-English language and serve as queries for gathering related news texts. Selected RSS feeds from under-resourced languages are also used for the same reasons. Documents are paired across monolingual collections by using a number of features (e.g., date and time of publication and similarity of title length and title content).

For **Wikipedia texts**, we developed a technique to find comparable Wikipedia texts based on the idea that inter-lingually linked Wikipedia text pairs containing significant numbers of shared anchor texts are likely to be quite similar in content (Paramita et al., 2012).

For **narrow domain texts**, a topic definition (specified as a list of topic terms) and a seed URL list are given to a focused monolingual crawler (FMC) that crawls starting from the seed URLs and performs lightweight text classification on pages it encounters to determine if they are relevant to the domain. For a specific topic, all of the returned texts in one language may be paired with all of the texts from another language to form a comparable corpus.

The above described tools are used to gather very large collections of comparable documents for all project language pairs. For news texts comparable corpora for 8 language pairs are collected, with the number of

document pairs ranging from 16,144 to 129,341. The "wikipedia-anchors" method contains corpora in 12 language pairs, with the number of document pairs ranging from 841 to 149,891. The FMC tool is used to collect narrow domain comparable corpora from the Web: 28 comparable corpora in 8 narrow domains for 6 language pairs have been constructed and amount to a total of more than 148M tokens.

### 3. Criteria of comparability and comparability metrics

In ACCURAT comparability is defined by how useful a pair of documents is for machine translation. Within the project, two different metrics are implemented to identify comparable documents from raw corpora crawled from the Web and to characterise the degree of their similarity (Su and Babych, 2012).

**The machine translation based metric** first uses the available machine translation API's for document translation and incorporates several useful features into the metric design. These features, including lexical information, keywords, document structure, and named entities, are then combined in an ensemble manner.

**The lexical mapping based metric** uses automatically generated GIZA++ bilingual dictionaries for lexical mapping. If a word in the source language occurs in the bilingual dictionary, the top 2 translation candidates are retrieved as possible translations in the target language. This metric provides a much faster lexical translation process, although word-for-word lexical mapping results are not as good as automatic translations.

The reliability of the proposed metrics has been tested on semi-manually collected Initial Comparable Corpora (Skadiņa et al., 2010b) used as a gold standard. It turned out that the comparability scores obtained from the comparability metrics reliably reflect comparability levels, as the average scores for higher comparable levels are always significantly larger than those of lower comparable levels. However, for the lexical mapping based metric, the average score for each comparability level drops in comparison to that of the MT based metric. The applicability of the proposed metrics was also measured by its impact on the task of parallel phrase extraction from comparable documents. The results show that a higher comparability level always leads to a significantly higher number of aligned phrases extracted from the comparable documents.

### 4. Alignment methods and information extraction from comparable corpora

In the ACCURAT project, the term alignment is used in the context of machine translation to describe the pairing of text in one document with its translation in another.

Through studies of existing alignment strategies designed for parallel corpora, comparable corpora, and non-comparable corpora, we showed that the most widely used alignment methods (Giza++ and Moses) are not well suited for use directly on strongly and weakly comparable texts. Therefore, the project consortium proposed new

methods and implemented tools that allow the alignment of comparable documents and the extraction of information (paragraphs, phrases, terminology, and named entities) from comparable corpora. All of the important tools that have been developed within the ACCURAT project for the alignment of comparable corpora at different levels and for data extraction from comparable corpora that are useful for machine translation are packed into the ACCURAT Toolkit (ACCURAT D2.6, 2011)<sup>1</sup>. By using the ACCURAT Toolkit, users may expect to obtain:

- **Comparable document (and other textual unit types) alignment.** This will facilitate the task of parallel phrase extraction by massively reducing the search space of such algorithms;
- **Parallel sentence/phrase mapping** from comparable corpora (Ion, 2012). This aims to supply clean parallel data useful for statistical translation model learning;
- **Translation dictionaries** extracted from comparable corpora. These dictionaries are expected to supplement existing translation lexicons which are useful for both statistical and rule-based MT;
- **Translated terminology** extracted (mapped) from comparable corpora (Ștefănescu, 2012). This type of data is presented in a dictionary-like format and is expected to improve domain-dependent translation;
- **Translated named entities** extracted (mapped) from comparable corpora. Also presented in a dictionary-like format, these lexicons are expected to improve parallel phrase extraction algorithms from comparable corpora and be useful by themselves when actually used in translation.

In order to map terms and named entities bilingually, the ACCURAT Toolkit also provides tools for detecting and annotating these types of expressions in a monolingual fashion.

The tools can be applied individually or in the provided workflows: (1) for parallel data mining from comparable corpora and (2) for named entity/terminology extraction and mapping from comparable corpora.

### 5. Comparable corpora in MT systems

With the ACCURAT toolkit, the consortium is aligning comparable corpora collected from the Web at the document level and extracting MT-related data – parallel phrases/sentences and bilingual lists of named entities and terminology. To evaluate the efficiency and usability of the developed methods for under-resourced languages and narrow domains, data extracted using these methods is being integrated into ACCURAT baseline MT systems. ACCURAT baseline MT systems are built for 17 translation routes using existing SMT techniques on available parallel corpora, e.g., JRC-ACQUIS

<sup>1</sup> <http://www accurat-project.eu/index.php?p=toolkit>

Multilingual Parallel corpus and SETimes corpus are available on MT-Serverland software infrastructure<sup>2</sup> and via the Web service.

For narrow domain MT, several successful proof-of-concept experiments were carried out to show that even small amounts of parallel domain specific data will help improve a SMT system.

To test the quality and effect of the data extracted with ACCURAT tools, an experiment with English-German domain-adapted SMT was performed for the automotive industry domain (Ștefănescu et al., 2012). By adding 45,952 sentence pairs extracted from the automotive domain comparable corpus, approximately 6.5 BLEU points over the baseline system were obtained.

The language model adaptation experiment was applied to the renewable energy domain. This led to improvements in terms of BLEU score for the following language pairs: for English->Greek BLEU increased from 15.07 to 15.14, for English->Lithuanian from 18.23 to 23.38, and for English->Croatian from 11.93 to 14.94.

More experiments are in progress, as corpora collection was finished recently and data extraction is still in progress.

## 6. Conclusion

The project is in its final phase now. The following key methods and tools are developed: crawling methods to identify comparable documents on the Web; comparability metric allowing identify comparable documents and evaluate their similarity; methods for automatic extraction of parallel and quasi-parallel data from any degree of comparable corpora.

## 7. Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347.

## 8. References

- ACCURAT D2.6 (2011) Toolkit for multi-level alignment and information extraction from comparable corpora., 31<sup>st</sup> August 2011 (<http://www accurat-project.eu/>), 123 pages.
- Aker, A.; Kanoulas, E. and Gaizauskas, R. (2012) A light way to collect comparable corpora from the Web. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.
- Eisele, A. and Xu, J. (2010). Improving machine translation performance using comparable corpora. In: *Proceedings of 3rd Workshop on Building and Using Comparable Corpora*, Malta, pp. 35--41.
- Ion, R. (2012) PEXACC: A Parallel Data Mining Algorithm from Comparable Corpora. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.
- Skadiņa, I.; Vasiļjevs, A.; Skadiņš, R.; Gaizauskas, R.; Tufiș, D., Gornostay, T. (2010a). Analysis and Evaluation of Comparable Corpora for Under

-Resourced Areas of Machine Translation. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, European Language Resources Association (ELRA), La Valletta, Malta, May 2010, pp. 6--14.

Skadiņa, I.; Aker, A.; Giouli, V.; Tufis, D.; Gaizauskas, R.; Mieriņa M. and Mastropavlos, N. A. (2010b). Collection of Comparable Corpora for Under-resourced Languages. In *Proceedings of the Fourth International Conference Baltic HLT 2010*, IOS Press, Frontiers in Artificial Intelligence and Applications, Vol. 219, pp. 161--168.

Skadiņa, I.; Aker, A.; Mastropavlos, N.; Su, F.; Tufis, D.; Verlic, M.; Vasiļjevs, A.; Babych, B.; Clough, P.; Gaizauskas, R.; Glaros, N.; Paramita, M.; Pinnis, M. (2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In *Proceedings of LREC 2012*, 21-27 May, Istanbul, Turkey.

Ștefănescu, D. (2012). Mining for Term Translations in Comparable Corpora. In *Proceedings of BUCC 2012*, May, 26, Istanbul, Turkey.

Ștefănescu, D.; Ion, R., and Hunsicker, S. (2012). Hybrid Parallel Sentence Mining from Comparable Corpora. In *Proceedings of EAMT 2012*.

<sup>2</sup> <http://www.dfki.de/mt-serverland/>