New Horizons in Translation and Interpreting Studies (Full papers) • December 2015 • pp.133-141

Towards a Web-based Tool to Semi-automatically Compile, Manage and Explore Comparable and Parallel Corpora

Hernani Costa[a], Gloria Corpas Pastor[a],
Ruslan Mitkov[b] and Miriam Seghiri[a]

[a]LEXYTRAD, University of Malaga, Spain
[b]RIILP, University of Wolverhampton, UK
{hercos,gcorpas}@uma.es, r.mitkov@wlv.ac.uk, seghiri@uma.es

Abstract

This article presents an ongoing project that which aims to design and develop a robust and agile web-based application capable of semi-automatically compiling multilingual comparable and parallel corpora, named iCorpora. Its main purpose is to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora. iCorpora intends to fulfil not only translators' and interpreters' needs, but also the needs of other professionals and laypeople, either by solving some of the usability problems found in the current compilation tools available on the market or by reducing their limitations and performance issues.

1 Introduction

In the last decade, there has been a growing interest in bilingual and multilingual corpora. In translation, in particular, their benefits have been demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volumes of data are just some of their advantages. Thus, it is not surprising that the use of corpora has been considered an essential resource in several research domains such as translation, terminology, language teaching, and automatic and assisted translation, amongst others. In particular, parallel corpora have become a very important source of knowledge, especially for Machine Translation (MT). Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) are just some examples of MT sub-areas where this kind of resource is fundamental, e.g. for the process of training (Hutchins and Somers, 1992). Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement in these areas. One potential solution to the insufficient parallel translation data is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora – i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES, 1996; Corpas Pastor, 2001:158). Although comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translation quality for under-resourced languages and narrow domains, the problem of data collection is still a significant technical challenge.

Bearing this in mind, the iCorpora project (cf. Costa et al., 2014c; 2015) proposes not

only to create a user-friendly interface to compile parallel corpora, but also to exploit comparable corpora from the Web. Broadly speaking, this ambitious project aims to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora by creating a new web-based application from scratch.

2 Existing Corpora Compilation Tools

The World Wide Web has become a primary meeting place for information and recreation, for communication and commerce. Millions of users have created billions of webpages in which they expressed their views about the world. As a source of machine-readable texts for corpus linguists and researchers in related fields such as Natural Language Processing (NLP) and MT for example, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. To this end, several tools (e.g. web crawlers, language identifiers, HTML parsers, HTML cleaners, etc.) have been developed and combined in order to produce corpora from this 'goldmine'. Therefore, this section aims to describe the most relevant approaches, methodologies, and tools capable of exploiting parallel and comparable corpora from the Web.

2.1 Mining Parallel Corpora

The Internet can be already considered a large multilingual corpus due to its huge number of multilingual websites, in which different pages can contain the same written text in different languages. This means that some of their webpages can be paired into *bitexts* (or parallel texts) – a very important source of knowledge, especially for MT systems. Nevertheless, the problem of collecting these data is still a significant technical challenge and the question remains: How can we find these parallel texts and obtain an aligned parallel corpus from them? Some attempts to answer this question are presented below.

STRAND[1] (Structural Translation Recognition, Acquiring Natural Data) (Resnik, 1998; 1999; Resnik and Smith, 2003) can be considered as one of the earliest core web-mining architectures capable of identifying webpages which are candidates to be bitexts. In order to do this, it uses the structural features of documents, a content-based measure of translational equivalence, and the Web as a source for mining bitexts on a large scale. The general procedure includes three main steps: 1) locate possibly parallel webpages; 2) generate candidate pairs of parallel webpages; and, finally, 3) apply structural filters to the candidate set. The details about the process can be found in Resnik, 1998; 1999; Resnik and Smith, 2003.

Bitextor[2,3] (Esplà Gomis, 2009; Esplà Gomis and Forcada, 2009; 2010) is a free/open-source application created for Unix platforms, which aims to generate translation memories using multilingual websites as a corpus source. This tool was created to be as adaptable as possible when retrieving multilingual data from any kind of website and work with any pairs of languages. To do that, it combines context-based and URL-based heuristics to harvest aligned *bitexts* from multilingual websites. The Bitextor workflow can be divided into three main steps: 1) downloading, processing and choosing the parameters for the comparison; 2) webpage comparison; and, finally, 3) aligning the obtained webpages. It is important to mention that Bitextor is based on two main assumptions: parallel pages should be under the same domain and they should have similar HTML structure.

Although this section only describes two systems, BITS (Ma and Liberman,

[1] `http://www.umiacs.umd.edu/~resnik/strand/`
[2] `http://bitextor.sourceforge.net`
[3] `http://sourceforge.net/projects/bitextor`

1999), PTMiner (Chen and Nie, 2000), WeBiTex[4] (Désilets et al., 2008) and ILSP-FC (Papavassiliou et al., 2013) should also be mentioned as they were developed for the same purposes.

2.2 Mining Comparable Corpora

There is a growing literature on using the Web for constructing various types of text collections, including domain-specific monolingual, bilingual and multilingual comparable corpora. Although the process of compiling comparable corpora can be manually performed, nowadays specialised tools can be used to automate this tedious task. This section presents the two best-known tools on the market for exploiting corpora mined from the Web.

BooTCaT[5] (Baroni and Bernardini, 2004) is a free and open-source semi-automatic compilation application that makes use of online information to construct web-based corpora. The process is very simple and only requires a set of seed terms as input. Then, these seeds are randomly grouped to form tuples (i.e. a variety of combinations of the seeds), which are submitted as search query strings to a search engine. It is possible to build a larger corpus by repeating the process using more seeds, or even create a comparable corpus by repeating the process using translational equivalents. Despite the multiple advantages, BootCaT has a few limitations, which restricts the "natural process" that is usually used to compile bilingual or multilingual comparable corpora (cf. Baroni and Bernardini, 2004:1313 and Gutiérrez Florido et al., 2013:3).

WebBootCat[6] (Baroni et al., 2006) is similar to BootCaT, but instead of having to download and install the application, WebBootCat can be used online. Yet, it is only freely available on a trial basis or through subscription.

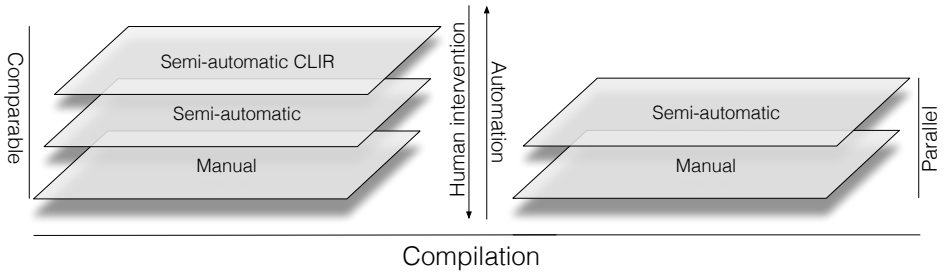
Although designed for other purposes, Terminus[7] and Corpográfo[8] should also be mentioned as examples of web-based compilation tools.

3 iCorpora: Compiling, Managing and Exploring Multilingual Data

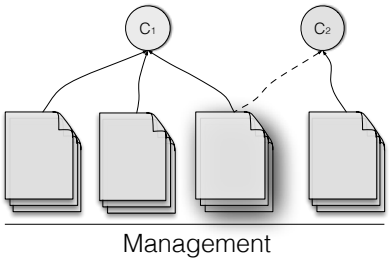
As shown in the previous section, several semi-automatic compilation tools have been proposed so far, capable of exploiting either comparable or parallel corpora from the Web. However, these compilation tools are sometimes scarce, proprietary, simplistic with limited features or too complex to be used by laypeople. Moreover, comparable compilation tools were built to compile one monolingual corpus at a time and do not cover the entire compilation process (i.e. apart from compiling monolingual comparable corpora, they do not allow the managing and exploration of both parallel and multilingual comparable corpora). Thus, their simplicity, lack of features, performance issues and usability problems result in a pressing need to design new compilation tools tailored to fulfil not only translators' and interpreters' needs (cf. Costa et al. (2014b;a)), but also the needs of professionals and laypeople.

After a careful analysis of the shortcomings and strengths of the current compilation tools, we started designing and developing a robust and agile web-based application prototype to semi-automatically compile, manage and explore both parallel and multilingual comparable corpora, which we named *iCorpora*. In detail, iCorpora will aggregate three applications: *iCompileCorpora*, *iManageCorpora* and *iExploreCorpora*.

[4] http://www.webitext.com
[5] http://bootcat.sslmit.unibo.it
[6] https://www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat
[7] http://terminus.iula.upf.edu//cgi-bin/terminus2.0/terminus.pl
[8] http://www.linguateca.pt/corpografo/



(a) *iCompileCorpora.*



(b) *iManageCorpora.*

(c) *iExploreCorpora.*

Figure 1: *iCorpora layered model.*

3.1 iCompileCorpora

iCompileCorpora can be simply described as a web graphical interface which will guide the user through the entire corpus compilation process. It will not only provide a simple interface with easy-to-follow steps, but will also enable experienced users to set advanced compilation options during the process.

3.1.1 Compiling Comparable Corpora

The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer (Figure 1a). This design option will not only result in increase of the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer. Specifically, the manual layer represents the option of compiling monolingual and multilingual comparable corpora, and will enable the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of both mono- and multilingual comparable corpora mined from the Internet. Although this layer can be considered similar to the approaches used by BootCaT and WebBootCat (see section 2.2), it has been designed to address some of their limitations (e.g. by allowing the use of more than one Boolean operator when creating search query strings). As there is now an increasing demand for systems that can somehow cross the language boundaries by retrieving information in various languages with just one query, the third layer aims to meet this demand by taking advantage of CLIR techniques to find relevant information written in a language different to the one semi-automatically retrieved by the methodology used in the previous layer.

3.1.2 Compiling Parallel Corpora

Regarding the parallel compilation process, iCompileCorpora will also facilitate for the manual upload of parallel documents from a local or remote directory onto the platform

(Figure 1a, manual layer). The second layer, i.e. the semi-automatic layer will offer the option of exploring parallel corpora mined from the Web. As shown in section 2.1, acquiring parallel data involves several tasks, such as crawling the web, parsing the structure of each fetched webpage and extracting its metadata, cleaning, classifying text, identifying near-duplicates, etc. Bearing this in mind, efficient focused web crawlers can be built by adapting existing open-source frameworks like Heritrix[9], Nutch[10] and Bixo[11]. Search engine Application Programming Interfaces (APIs) can also be used to identify in-domain webpages (Hong et al., 2010) or multilingual web sites (Resnik and Smith, 2003). At this point it is not yet clear which approach/algorithms and/or frameworks iCompileCorpora will use. Nevertheless, the methodology proposed in Resnik, 1998; 1999; Resnik and Smith, 2003 seems to be the most commonly used, i.e. locate possibly parallel webpages, generate candidates pairs of parallel webpages, and then apply structural filters to the candidate set in order to clean "noisy data".

3.2 iManageCorpora

The second application is called iManageCorpora (Figure 1b). This application will be specially designed to: manage (i.e. make it possible to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and explore the representativeness of the corpora (cf. Corpas Pastor and Seghiri, 2009).

3.3 iExploreCorpora

Finally, iExploreCorpora (Fig. 1c) intends to offer a set of concordance features, such as the ability to search for words in context and automatically extract the most frequent words and multiword units, amongst other features.

[9] http://crawler.archive.org/
[10] `http://nutch.apache.org`
[11] `http://openbixo.org/`

4 Concluding Remarks

Against the background of the increasing importance of multilingual data, iCorpora's objectives are to develop a novel, flexible and robust web-based application for the compilation, management and exploitation of comparable and parallel corpora and to address the needs of translators and interpreters as well as other professional and casual users. This ongoing project aims to increase the flexibility and robustness of the compilation process by solving some of the usability problems found in the current compilation tools available on the market or by reducing their limitations and performance issues. By the end of this project, we intend to make this compilation tool publicly available, both in a research and in a commercial setting.

Acknowledgements

Hernani Costa is supported by the People Programme (Marie Curie Actions) of the European Union's Framework Programme (FP7/2007-2013) under REA grant agreement n 317471. The research reported in this work has also been partially carried out in the framework of the Educational Innovation Project TRADICOR (PIE 13-054, 2014-2015); the R&D project INTELITERM (ref. n FFI2012-38881, 2012-2015); the R&D Project for Excellence TERMITUR (ref. n HUM2754, 2014-2017); and the LATEST project (ref. 327197-FP7-PEOPLE-2012-IEF). We would also like to thank Emma Franklin for proof-reading this paper.

References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *6th Conf. on Applied Natural Language Processing*, pages 21–28.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Studien Zur Romanischen Sprachwissenschaft Und Interkulturellen Kommunikation, 49. Peter Lang Pub Incorporated, Frankfurt, Germany.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014a). A comparative User Evaluation of Terminology Management Tools for Interpreters. In *25th Int. Conf. on Computational Linguistics (COLING'14), 4th Int. Workshop on Computational Terminology (CompuTerm'14)*, pages 68–76, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Costa, H., Corpas Pastor, G., and Durán Muñoz, I. (2014b). Technology-assisted Interpreting. *MultiLingual #143*, 25(3):27–32.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014c). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). iCorpora: Compiling, Managing and Exploring Multilingual Data. In *7th Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. (2008). WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content. In *Translating and the Computer 30*, London, UK.

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

Esplà Gomis, M. (2009). Bitextor, un cosechador automático de memorias de traducción a partir de sitios web multilingües. *Procesamiento del Lenguaje Natural*, 43(1):365–366.

Esplà Gomis, M. and Forcada, M. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. In *Workshop Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.

Esplà Gomis, M. and Forcada, M. (2010). Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(1):77–86.

Gutiérrez Florido, R., Corpas Pastor, G., and Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In *10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounters*, Paris, France.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). An Empirical Study on Web Mining of Parallel Data. In *23rd Int. Conf. on Computational Linguistics*, COLING'10, pages 474–482. ACL.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Ma, X. and Liberman, M. (1999). BITS: A Method for Bilingual Text Search over the Web. In *Machine Translation Summit VII*.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *6th Workshop on Building and Using Comparable Corpora (BUCC'13)*, pages 43–51, Sofia, Bulgaria. ACL.

Resnik, P. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. In *3rd Conf. of the Association for Machine Translation in the Americas*, AMTA'98, Langhorne, PA, USA. Lecture Notes in Artificial Intelligence 1529.

Resnik, P. (1999). Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL'99, pages 527–534. ACL.

Resnik, P. and Smith, N. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.