SmartBiC: Smart Harvesting of Bilingual Corpora from the Internet

Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Alicia Núñez Alcover, Tudor N. Mateiu, Mikel L. Forcada
Prompsit Language Engineering
info@prompsit.com

Pedro L. Díez Orzas, Almudena Ballester Carrillo, Giuseppe Deriard Nolasco, Noelia Jiménez Listón
Linguaserve Int. de Servicios
clientes@linguaserve.com

Abstract

SmartBiC, an 18-month innovation project funded by the Spanish Government, aims at improving the full process of collecting, filtering and selecting in-domain parallel content to be used for machine translation and language model tuning purposes in industrial settings. Based on state-of-the-art technology in the free/open-source parallel web corpora harvester Bitextor, SmartBic develops a web-based application around it including novel components such as a language -and domain- focused crawler and a domain-specific corpora selector. SmartBic also addresses specific industrial use cases for individual components of the Bitextor pipeline, such as parallel data cleaning. Relevant improvements to the current Bitextor pipeline will be publicly released.

1 Introduction

Obtaining a suitable amount of parallel corpora to train neural machine translation (NMT) or large language models (LLMs) is a challenging task, which becomes particularly difficult for domain-specific areas. This is seen as a severe limitation for the full development of NMT and LLMs in industrial settings. SmartBic, an innovation project ending in September 2023 with support from the Spanish Government through the NextGenerationEU funds, addresses this limitation from an industrial perspective. The main goal of SmartBiC is to ease the collection, filtering and selection of in-domain parallel corpora by exploring multilingual websites and external corpora. To that end, SmartBiC builds upon Bitextor,[1] a free/open-source parallel corpora harvester, adding the following improved and novel components described in detail in section 3: a smart crawler, a smart cleaner, and a smart corpora selector. In-domain corpora produced with SmartBiC will be used in industrial use cases, particularly to train domain-specific NMT systems, but also to explore their usefulness as LLM training or tuning datasets. This will extrinsically evaluate the quality and usefulness of the results of the project and point to improvements for future work.

2 Bitextor as the core technology

Bitextor (Espla-Gomis et al., 2016) has been used to produce petabytes of parallel corpora from multilingual web-crawled content in previous EU-funded projects such as ParaCrawl.eu or MaCoCu.eu. Among other alternatives, such as the ILSP-FC focused crawler (Papavassiliou et al., 2013), Bitextor is chosen for its language support, active developer community and modular design.

3 New components in SmartBiC

SmartBiC addresses the following limitations in Bitextor in order to achieve the goals of the project: (1) crawlers in Bitextor download websites ignoring the language pair indicated in the input, only a tiny portion of the crawled content makes it to the output, (2) no topic or domain constraints are currently handled by Bitextor, in-domain and generic content is mixed in the output parallel corpus and (3) seed URLs need to be manually provided to the crawler and no mechanism to automatically discover further interesting URLs is available.

[1] https://github.com/bitextor/bitextor

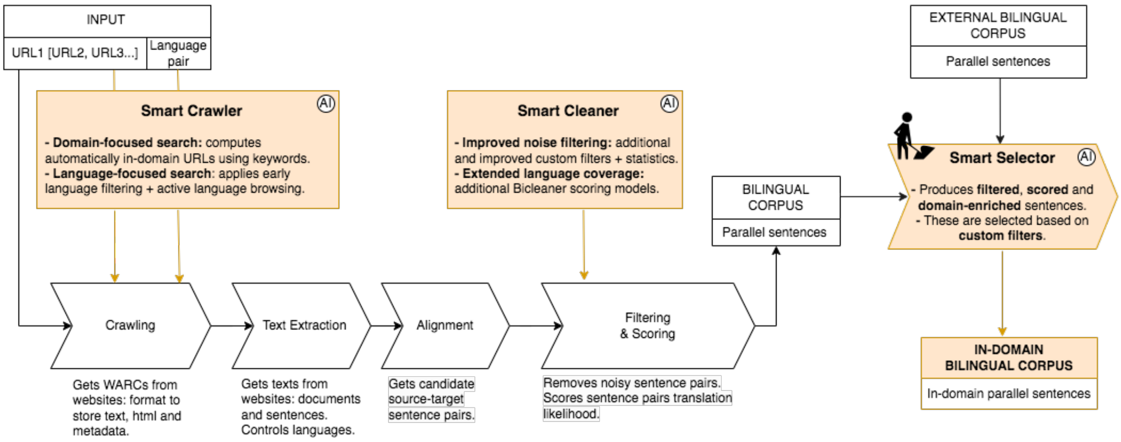
INPUT
URL1 [URL2, URL3...]
Language pair
Smart Crawler
AI
- Domain-focused search: computes automatically in-domain URLs using keywords.
- Language-focused search: applies early language filtering + active language browsing.
Smart Cleaner
AI
- Improved noise filtering: additional and improved custom filters + statistics.
- Extended language coverage: additional Bicleaner scoring models.
EXTERNAL BILINGUAL CORPUS
Parallel sentences
Smart Selector
AI
- Produces filtered, scored and domain-enriched sentences.
- These are selected based on custom filters.
BILINGUAL CORPUS
Parallel sentences
Crawling
Text Extraction
Alignment
Filtering & Scoring
IN-DOMAIN BILINGUAL CORPUS
In-domain parallel sentences
Gets WARCs from websites: format to store text, html and metadata.
Gets texts from websites: documents and sentences. Controls languages.
Gets candidate source-target sentence pairs.
Removes noisy sentence pairs. Scores sentence pairs translation likelihood.

Figure 1: Bitextor pipeline (uncoloured boxes) and new and enhanced components provided by SmartBiC (coloured boxes).

SmartBiC adds as new components to Bitextor to overcome these limitations (see Figure 1):

The **Smart Crawler** module which improves the current generic crawler by integrating both language pair and domain focus during crawling and includes: (1) two new crawlers with complementary crawling strategies which filter content by language pair, reducing the amount of information downloaded; (2) a mechanism to automatically find URLs from another URL, a text or a keyword list Exploiting these keywords, potentially similar URLs are discovered using browser-based queries.

The **Smart Cleaner**, a stand-alone component that enhances and provides the functionality of the Bitextor cleaning pipeline; it includes: (1) improved and new filters with customisable thresholds; (2) support for new language pairs, that is, new models for Bicleaner; (3) custom filtering, cleaning statistics and new output formats.

The **Smart Selector** independent module which picks the most relevant set of in-domain data from already crawled or generic corpora and includes: (1) a separate, standalone cleaning step to filter out noisy sentence pairs and to score the remaining ones with Bicleaner (Zaragoza-Bernabeu et al., 2022); (2) a multilingual zero-shot classifier which adds domain scores to each sentence in a sentence pair; (3) a customisable sentence pair selector based on size, domain, cleaning scores and content quality.

Relevant components resulting from this project will be released through Bitextor.

3.1 SmartBiC web application

SmartBiC will be operated mainly by computational linguists and translators through a web application developed within the project. The main features of this web application, which adds usability to current command-line-only Bitextor, will allow users to launch and monitor crawling, cleaning and in-domain data selection tasks.

3.2 Acknowledgements

SmartBiC is funded by NextGenerationEU funds of the Spanish Government through the grants for Artificial Intelligence Research and Development projects and other digital technologies and their implementation in value chains (C005/21-ED) by “Entidad Pública Empresarial RED.ES, M.P.”, grant number 2021/C005/00150077.

References

Espla-Gomis, Miquel, Mikel L Forcada, Sergio Ortiz-Rojas, and Jorge Ferrández-Tordera. 2016. Bitextor’s participation in wmt’16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 685–691.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor Thurmair. 2013. A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria, August. Association for Computational Linguistics.

Zaragoza-Bernabeu, Jaume, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France, June. European Language Resources Association.