Discovering Parallel Language Resources for Training MT Engines

Vassilis Papavassiliou, Prokopis Prokopidis, Stelios Piperidis
Institute for Language and Speech Processing
Athena Research and Innovation Center, Athens, Greece
{vpapa, prokopis, spip}@ilsp.gr

Abstract

Web crawling is an efficient way for compiling the monolingual, parallel and/or domain-specific corpora needed for machine translation and other HLT applications. These corpora can be automatically processed to generate second order or synthesized derivative resources, including bilingual (general or domain-specific) lexica and terminology lists. In this submission, we discuss the architecture and use of the ILSP Focused Crawler (ILSP-FC), a system developed by researchers of the ILSP/Athena RIC for the acquisition of such resources, and currently being used through the European Language Resource Coordination effort. ELRC aims to identify and gather language and translation data relevant to public services and governmental institutions across 30 European countries participating in the Connecting Europe Facility (CEF).

Keywords: parallel language resources, document pair detection, web crawling, machine translation

1. ILSP-FC Architecture

ILSP-FC[1] is a comprehensive solution for acquiring domain-specific, monolingual and bilingual datasets from the web. One of the main components of the system is an efficient crawler that initializes its frontier (i.e. the list of pages to be visited) from a seed URL list provided by the user, classifies fetched pages as appropriate for the user's aims (i.e. in the targeted language and/or relevant to the targeted domain), extracts links from fetched web pages, adds them to the list of pages to be visited and repeats this process until an expiration criterion is reached. In the case of focused crawls for domain-specific content, the input expected from the user also includes a domain profile, i.e. a list of terms that describe the domain.

In order to ensure scalability, the system is based on open-source libraries that allow configuration of workflows that can be executed on top of the Hadoop framework for distributed data processing. Due to its modular architecture, each of the system's components can be easily substituted by alternatives with the same functionalities. The main components integrated in ILSP-FC (Papavassiliou et al., 2013) are:

Page Fetcher adopts a multi-threaded crawling implementation in order to ensure concurrent visiting of multiple web pages/hosts and fetching of user-targeted specific document types (e.g. html, docx, pdf).

Normalizer detects the text encoding of the downloaded web pages and if needed, converts it to UTF-8. It also parses the structure of each web page and extracts its metadata (e.g. title, description, keywords, publisher, author, license etc.). In order to extract textual content and metadata from a set of formats (txt, docx and pdf), it uses open libraries[2].

Cleaner segments the main text in paragraphs, identifies boilerplate (e.g. advertisements, disclaimers) and extracts structural information like titles, headings and list items. For this task, a modified version of Boilerpipe (Kohlschütter et al., 2010) is used.

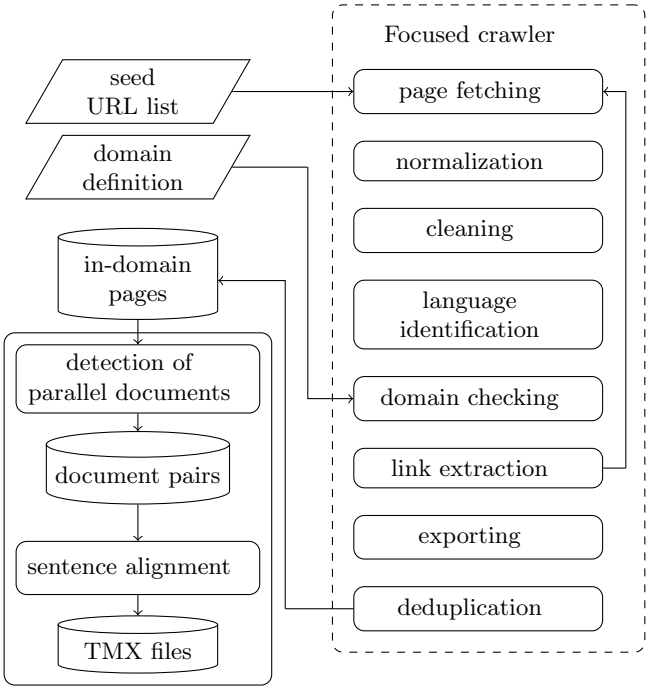

Figure 1: ILSP-FC workflow for acquisition of parallel LRs

Language Identifier detects the language of a document, as well as paragraphs in a language different from the main one. It uses open source language identification libraries like `lang-detect`[3], that perform at over 99% precision at document level for more than 50 languages.

Domain checker compares the content of the page to a user-provided domain profile. Based on the number of terms' occurrences, their location in the web page and the weights of found terms, a page relevance score is calculated. This score is compared with a predefined threshold and the web page is categorized as relevant to the domain.

Link Extractor examines the "neighbourhood" of links for terms, special patterns and/or semantic annotations (e.g.

[1] `http://nlp.ilsp.gr/ilsp-fc/`

[2] `https://pdfbox.apache.org/`, `https://poi.apache.org/text-extraction.html`

[3] `https://github.com/shuyo/language-detection`

link[hreflang], a[id=linkTranslateVersion], etc.), and ranks links by the probability that they point to candidate translations and/or in domain pages. Thus it organizes the list of URLs to be fetched as a priority queue and guides the crawler to visit "interesting" pages first.

De-duplicator checks each document against all others and identifies (near) duplicates based on lists of quantised word frequencies extracted from each document; and on the percentage of common paragraphs (e.g. 80%). In case a pair of (near) duplicates has been detected, the shortest is discarded.

Pair Detector (a.k.a. bitext identification module) identifies documents that could be considered parallel. This module exploits the results of online discovery and prioritization of translation links during crawling, and uses alternative criteria based on special patterns included in the URLs, cooccurrences of images with the same filename, similar sequences of digits, and structure similarity. It does not use any knowledge of the targeted languages (e.g. bilingual lexica or MT output) but applies language independent methods for pair detection.

Sentence Aligner uses open-source aligners (e.g. Hunalign (Varga et al., 2005), Maligna (Jassem and Lipski, 2008)) to extract sentence alignments from document pairs, and exports results in TMX (i.e. a TMX file for each identified document pair).

As an alternative to the pipeline use of the tool for acquisition of parallel LRs from the web (cf. Figure 1), specific modules (e.g. domain checking, document pairing, and sentence alignment) can be called as standalone modules for all relative tasks. They can therefore be used for the processing of resources residing in proprietary data repositories. ILSP-FC is available under a GPL license. Licensing and support alternatives for commercial uses and applications are also available.

2. Adapting to ELRC requirements for parallel LRs

The European Language Resource Coordination[4] (Lösch et al., 2018) effort aims to identify and gather language and translation data relevant to public services and governmental institutions across 30 European countries participating in the Connecting Europe Facility[5]. Through a series of workshops, ELRC has been showcasing the benefits of the CEF automated translation platform (CEF eTranslation) and has been trying to mobilize public sector representatives to share public language resources. These LRs will contribute in enhancing CEF eTranslation and, in the end, providing EU citizens with better multilingual services. To complement these data gathering efforts, ELRC has also employed ILSP-FC and other automatic methods for acquiring multilingual LRs. In the course of the project and based on feedback by all consortium partners, the tool was continuously tested and enhanced at ILSP in order to provide more accurate results. It was eventually deployed at all four ELRC partner sites for acquiring language resources for specific (EN-X) language pairs, where X stands for official EU languages in CEF-affiliated countries[6].

In order to meet ELRC needs to cover all CEF languages, missing resources were constructed and integrated in the tool (for example language profiles for both Norwegian written standards, Bokmål and Nynorsk). The accuracy of the language identifier, when examining Norwegian texts is 98%/90% for text chunks of at least 500/100 characters, respectively.

Turning to the identification of bitexts, the tool employs a combination of methods that are language-pair agnostic, i.e. they do not use bilingual lexica or MT results that are often difficult to generate. For evaluation purposes, the bitext identification module was submitted (Papavassiliou et al., 2016) in the WMT 2016 Bilingual Document Alignment Shared Task (Buck and Koehn, 2016) and scored a high recall of 91%. It ranked 7th among 21 participations, and, to the best of our knowledge, first among those not using language-pair specific resources or MT output as a feature for document alignment.

For segment alignment, the system uses open source aligners to construct collections of candidate parallel segments. A battery of criteria are applied on these candidates with the purpose of filtering out or annotating automatically specific types of translation units (TUs) (i.e. detecting potential alignment or translation issues, or sentence pairs of limited or no use) and of generating precision-high multilingual LRs for training MT systems. This filtering was carried out by:

(i) identifying duplicate TUs, or TUs with identical text in both languages,

(ii) estimating the alignment quality by calculating the so-called alignment score,

(iii) excluding TUs based on the segment length ratio (note: segments with a length ratio close to 1 have similar length, whereas segments having a ratio far from 1 would indicate potential alignment problems)

(iv) identifying TUs in which numbers in the segment for one language are different compared to the segment for the other language,

(v) excluding TUs that are included in document pairs that contain many TUs (e.g. over 40%) of type 0:1, an indication that such document pairs consist of comparable rather than parallel documents,

(vi) excluding TUs consisting solely of URLs, emails, and dates.

[4] http://lr-coordination.eu

[5] https://ec.europa.eu/inea/en/connecting-europe-facility

[6] Responsibility for data collection was shared among all ELRC consortium members in the following way: DFKI (6 countries): Germany, Austria, Luxembourg, Netherlands, Hungary, Czech Republic; ELDA (8 countries): Ireland, Spain, Portugal, Belgium, Italy, Malta, France, (U.K.); Tilde (8 countries): Latvia, Estonia, Lithuania, Finland, Sweden, Denmark, Iceland, Norway; ILSP (8 countries): Greece, Cyprus, Slovakia, Slovenia, Bulgaria, Poland, Romania, Croatia

lr-id	lr-name	"good"	"bad"
115	Parallel corpus (Greek - English) in the public administration domain	12332 (98.59%)	177 (1.41%)
379	Parallel corpus (Bulgarian - English) in the public administration domain	11094 (98.51%)	168 (1.49%)

Table 1: Automatic evaluation of parallel sentences contained in two LRs generated with ILSP-FC

The optional domain checking procedure during crawling can be complemented by the use of post-crawling topic classification with tools like the JRC Eurovoc Indexer (JEX (Steinberger et al., 2012)). We have used JEX to tag multilingual documents with identifiers that correspond to domains, micro-thesauri and thesauri concepts from Eurovoc[7], the EU's multilingual thesaurus.

3. Construction of parallel LRs out of public sector data

One of the methods employed for the construction of parallel LRs in the ELRC project, was to identify public administration websites (e.g. websites of ministries, local authorities, embassies, courts, etc.) as candidate sources for extracting content relevant to the CEF Digital Service Infrastructures (DSIs)[8]. Then ELRC consortium partners used ILSP-FC to crawl these websites and process their content in order to construct a mono/bilingual collection for each website. Finally, the outcomes of the websites were merged based on the language and the relevance of their content. For instance, the acquired content from websites of Polish cultural organizations was merged to generate a monolingual LR of 10.2M tokens and a parallel EN-PL LR of 36.3K TUs.

Although crawling with the tool for the purposes of ELRC is ongoing work, parallel LRs for several language pairs have already been generated and a number of them (EN-EL, EN-BG, EN-MT, EN-SV, EN-IS, EN-ET, EN-ES) have become available through the ELRC repository[9] (Piperidis et al., 2018) by consortium members.

4. Evaluation experiments

Assessing the usefulness of a system that discovers, acquires and transforms bitexts from the web involves many different evaluation questions: Does the system identify most of the document pairs published on a web site? Are these document pairs as noise-free as possible? Are the aligned sentences extracted from the document pairs clean enough for training MT systems? Apart from our participation in the WMT 2016 shared task, we also conducted experiments covering a variety of language pairs, in order to evaluate both the acquisition procedure and the acquired resources, focusing on parallelness (at document and segment level) and domainess.

In an effort to evaluate recall for the pair detector module, we used parallel datasets crawled from the Global Voices group of websites (Prokopidis et al., 2016). In this dataset, documents are connected with specific links when one is the translation of the other. Since this is not the case for many multilingual web sites, during the evaluation of the pair detector we omitted the tool's methods that exploit special patterns in URLs and links that point to translations. Thus, we only used methods that are based on a) cooccurrences of images with the same filename in HTML source, b) edit distance of sequences of digits in the main content of webpages and c) structural similarity. We evaluated these methods in the task of reconstructing the English-Greek parallel collection, i.e. of identifying the 3581 document pairs for this language pair. The recall and precision rates were 68.56% and 92.50% respectively. The main reason for the relatively low document-level recall is that many document pairs consisting of very short documents were not identified. However, the recall at token level (i.e. the percentage of tokens retrieved from all translated sentences) was 91.18%, a fact that implies that "lost" bitexts contained less that 9% of the actual parallel content.

In another evaluation experiment, we automatically estimated the quality of the sentence-level parallelness of the LRs created with ILSP-FC. To this end, we trained the C-Eval (Zariņa et al., 2015) parallel corpora cleaning and evaluation tool on the DGT-TM 2015 release[10] to build an automatic classifier identifying non-parallel sentences in a parallel corpus. We then classified the parallel sentences contained in two datasets delivered as ELRC LRs #115 (EN-EL) and #379 (EN-BG). Results in Table 1 indicate that LRs created with the tool include a high percentage ($\approx$ 98.5% for these specific LRs) of useful translation segments.

In order to test domainess, we used JEX to assign Eurovoc identifiers on the English text of LRs #115 and #379. In the results shown in Table 2, the assigned identifiers seem to correctly depict the "nature" of #155, for which about 47% and 29% of its content was acquired from the websites of the Greek Ministry of National Defence and the Ministry of Foreign Affairs respectively. Similarly, about 85% of the sentence pairs in #379 were compiled from the websites of the Ministry of Foreign Affairs, and the President of the Republic of Bulgaria. We also examined two parallel collections (EN-HR and EN-PL) in the *culture* domain. In the results in Table 3 only one identifier (*approximation of laws*) seems irrelevant to the targeted domain.

[7] `http://eurovoc.europa.eu/`

[8] CEF DSIs include Online Dispute Resolution; Europeana; Open Data Portal; eJustice; and Electronic Exchange of Social Security Information

[9] `https://elrc-share.eu`

[10] `https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory`

id	label	weight
1182	Greece	0,324
5335	military intervention	0,189
3489	cooperation policy	0,186
2628	armed forces	0,184
5786	military personnel	0,183
2499	European defence policy	0,176

id	label	weight
5063	Bulgaria	0,461
3763	Romania	0,295
3489	cooperation policy	0,211
914	Eastern Europe	0,207
12	accession to the European Union	0,169
1474	EC agreement	0,164

Table 2: Eurovoc identifiers assigned automatically by JEX for LRs #115 (EN-EL, top) and #379 (EN-BG, bottom)

id	label	weight
2023	music	0,217
2543	Poland	0,19
2459	cultural policy	0,157
208	cultural cooperation	0,143
529	copyright	0,138
2897	approximation of laws	0,109

id	label	weight
5563	Croatia	0,199
208	cultural cooperation	0,151
2840	heritage protection	0,143
2459	cultural policy	0,142
3200	cultural relations	0,135
4877	cultural object	0,125

Table 3: Eurovoc identifiers assigned automatically by JEX for an EN-PL (top) and an EN-HR (bottom) LR in the culture domain

5. Acknowledgements

This work has been supported by the European Language Resource Coordination (ELRC), a service contract that operated under the EC's Connecting Europe Facility SMART 2014/1074 programme from April 2015 to April 2017. ELRC continues until December 2019 under SMART 2015/1091 LOT 2 "Language Resource coordination and collection with related legal and technical work" and SMART 2015/1091 LOT 3 "Acquisition of additional Language Resources and related refinement/processing services and their provision of the Language Resource Repository of CEF Automated Translation Platform". We would like to thank ELRC consortium members from DFKI, TILDE, and ELDA, who provided feedback, reported issues and used the system described in this paper to acquire parallel resources for the purposes of the contract. Access to the Okeanos cloud service[11] provided by the Greek Research & Technology Network, is greatly appreciated.

[11] https://okeanos.grnet.gr/

6. Bibliographical References

Buck, C. and Koehn, P. (2016). Findings of the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany.

Jassem, K. and Lipski, J. (2008). A new tool for the bilingual text aligning at the sentence level. In *Proceedings of Intelligent Information Systems Conference*, Zakopane, Poland.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 441–450, New York, USA.

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T., Schnur, E., Choukri, K., and van Genabith, J. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (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.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ILSP/ARC submission to the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 733–739, Berlin, Germany.

Piperidis, S., Labropoulou, P., Deligiannis, M., and Giagkou, M. (2018). Managing public sector data for multilingual applications development. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Prokopidis, P., Papavassiliou, V., and Piperidis, S. (2016). Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In *Proceedings of the 10th Language Resources and Evaluation Conference*, Portorož, Slovenia.

Steinberger, R., Ebrahim, M., and Turchi, M. (2012). JRC Eurovoc Indexer JEX - A freely available multi-label categorisation tool. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 798–805, Istanbul, Turkey.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Zariņa, I., Ņikiforovs, P., and Skadiņš, R. (2015). Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 185–192, Antalya, Turkey.