Comparing two acquisition systems for automatically building an English–Croatian parallel corpus from multilingual websites

Miquel Esplà-Gomis,[*] **Filip Klubička,**[†] **Nikola Ljubešić,**[†]
Sergio Ortiz-Rojas,[‡] **Vassilis Papavassiliou,**[§] **Prokopis Prokopidis**[§]

[*]University of Alacant, Alacant (Spain)
mespla@dlsi.ua.es

[†]University of Zagreb, Zagreb (Croatia)
fklubicka@ffzg.hr, nikola.ljubesic@ffzg.hr

[‡]Prompsit Language Engenering, Elx (Spain)
sergio@prompsit.com

[§]Institute for Language and Speech Processing, Athens (Greece)
vpapa@ilsp.gr, prokopis@ilsp.gr

Abstract

In this paper we compare two tools for automatically harvesting bitexts from multilingual websites: *bitextor* and *ILSP-FC*. We used both tools for crawling 21 multilingual websites from the tourism domain to build a domain-specific English–Croatian parallel corpus. Different settings were tried for both tools and 10,662 unique document pairs were obtained. A sample of about 10% of them was manually examined and the success rate was computed on the collection of pairs of documents detected by each setting. We compare the performance of the settings and the amount of different corpora detected by each setting. In addition, we describe the resource obtained, both by the settings and through the human evaluation, which has been released as a high-quality parallel corpus.

Keywords: bitext crawling, parallel corpora, Croatian

1. Introduction

Parallel corpora are a valuable source of cross-lingual knowledge, consisting of collections of text-fragment pairs, usually known as *bitexts* (Harris, 1988), which are mutual translations in different languages. These corpora have been shown to be a useful resource for a wide range of tasks in natural language processing (Melamed, 2001), such as cross-lingual information retrieval (Nie et al., 1999), cross-lingual textual entailment (Mehdad et al., 2011), or word-sense disambiguation (Diab and Resnik, 2002). However, it is in statistical machine translation (SMT) (Koehn, 2010) where the use of parallel corpora is more relevant. The proliferation of parallel-corpora-based methods has raised a growing interest on parallel corpora collection in the last decades.

Many sources of bitexts have been identified: parallel corpora have been built from legal texts, such as the Hansards corpus (Roukos et al., 1995) or the Europarl corpus (Koehn, 2005); translations of software interfaces and documentation, such as KDE4 and OpenOffice (Tiedemann, 2009); or news translated into different languages, such as the SE-Times corpus (Tiedemann, 2009), or the News Commentaries corpus (Bojar et al., 2013), etc.

One of the hugest sources of parallel corpora is the Internet, since there are many websites which are available in two or more languages. Many approaches have been therefore proposed for trying to exploit the Web as a parallel corpus. One of the most complex tasks involved in this problem is parallel document identification. Three main strategies can be found in the literature for parallel document identifica-tion in multilingual websites by exploiting:

- similarities in the URLs corresponding to web pages from a web site (Ma and Liberman, 1999; Nie et al., 1999; Resnik and Smith, 2003; Chen et al., 2004; Zhang et al., 2006; Désilets et al., 2008; Esplà-Gomis and Forcada, 2010; San Vicente and Manterola, 2012);

- parallelisms in the structure of HTML files (Nie et al., 1999; Resnik and Smith, 2003; Sin et al., 2005; Shi et al., 2006; Zhang et al., 2006; Désilets et al., 2008; Esplà-Gomis and Forcada, 2010; San Vicente and Manterola, 2012; Papavassiliou et al., 2013); and

- content-similarity techniques (mostly based on bag-of-words overlapping metrics) (Ma and Liberman, 1999; Chen et al., 2004; Zhang et al., 2006; Jiang et al., 2009; Utiyama et al., 2009; Yan et al., 2009; Hong et al., 2010; Sridhar et al., 2011; Antonova and Misyurev, 2011; Barbosa et al., 2012).

In addition to these strategies, other heuristics can be found in the bibliography, such as file size comparison, language markers in the HTML structure, mutual hyper-links between web pages, or images co-occurrence (Papavassiliou et al., 2013). It is usual to combine several of these methods in order to improve the performance.

In this work we use two tools from this bibliography, *ILSP-FC*[1] (Papavassiliou et al., 2013) and *bitextor*[2]

---

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

[2]http://sf.net/projects/bitextor

(Esplà-Gomis and Forcada, 2010), for harvesting English–Croatian parallel documents from a collection of 21 multilingual websites belonging to the tourism domain. In our experiments, we compare the success rate of these settings to detect parallel documents by manually checking a representative sample of the document pairs obtained by each of them. Additionally, we describe the parallel corpus obtained as a by-product of this evaluation.

1.1. Bitextor

Bitextor is a free/open-source tool for harvesting bitexts from multilingual websites. The newest version of bitextor (version 4.0) is a re-implementation of the tool described by Esplà-Gomis and Forcada (2010). In this version, the techniques based on URL similarity are replaced by new methods based on bag-of-words overlapping. Given a multilingual website and the pair of targeted languages (L_1, L_2) from which the parallel corpus has to be created, bitextor performs the following steps:

1. the website is completely downloaded by means of the tool *HTTrack*,[3] keeping only HTML documents;

2. downloaded documents are preprocessed with *Apache Tika*[4] and *boilerpipe*[5] (Kohlschütter et al., 2010) to normalise the HTML structure and remove boilerplates;

3. duplicate documents (regarding the text, not the structure) are removed, and the language of each file is detected with *LangID* (Lui and Baldwin, 2012),[6] keeping only those documents in L_1 or L_2;

4. bag-of-words overlapping metrics are used to choose a preliminary n-best candidates list for each document;

5. each n-best candidates list is re-ranked by using metrics based on the Levenshtein edit distance between the HTML structure of each pair of documents;

6. the most promising document pairs in the n-best candidates lists are aligned and *hunalign*[7] (Varga et al., 2005) is used to obtain an indicative score regarding the quality of the sentence-alignment between both documents.

1.2. ILSP-FC

ILSP-FC is a modular system that includes components and methods for all the tasks required to acquire domain-specific corpora from the Web. Depending on user-defined configuration, the crawler employs processing workflows for the creation of either monolingual corpora or bilingual collections (i.e. pairs of parallel documents acquired from multilingual web sites). The main modules integrated in ILSP-FC are:

1. page fetcher: adopts a multithreaded crawling implementation in order to ensure concurrent visiting of multiple web pages/hosts.

2. normaliser: parses the structure of each fetched web page and extracts its metadata, detects its encoding and converts it to UTF-8 if required.

3. cleaner: extracts structural information (i.e. title, heading, etc.) and identifies boileplate paragraphs.

4. language identifier: uses the *Cybozu*[8] library to detect the main language of a document, as well as paragraphs in a language different from the main one.

5. link extractor: examines the anchor text of the extracted links and ranks them by the probability that a link from a page points to a candidate translation of this page, with the purpose of forcing the crawler to visit candidate translations first.

6. de-duplicator: checks each document against all others and identifies (near-)duplicates by comparing the quantized word frequencies and the paragraphs of each pair of candidate duplicate documents;

7. pair detector: examines each document against all others and identifies pairs of documents that could be considered parallel. Its main methods are based on URL similarity, co-occurrences of images with the same filename in two documents, and the documents' structural similarity.

2. Experimental settings

Our English–Croatian corpus is built from the collection of 21 multilingual websites listed in Table 1. These websites were handpicked from a list of 100 most bitext-productive multilingual websites from the Croatian top-level domain. The list of the most productive multilingual websites was obtained by calculating the website frequency distribution in the *hrenWaC* corpus[9] (Tiedemann, 2009), a side-product of the *hrWaC* Croatian web corpus (Ljubešić and Erjavec, 2011). Our future plans cover combining the procedure of top-level domain crawling for bitext-hotspot identification and bilingual focused crawling of the bitext hotspots for obtaining parallel data.

In our experiments, two different configurations were tried for ILSP-FC:

- *all*: It includes all the pairs detected by the tool (i.e. default configuration);

- *reliable*: It includes a subset of the *all* configuration where only those pairs identified through image co-occurrences and high-structural similarity are kept;

and four were tried for bitextor:

- *10-best*: 10-best candidate lists are used to get the pairs of documents;

- *1-best*: 1-best candidate lists are used to get the pairs of documents; this setting is more strict than *10-best*, since it only aligns documents which are mutual best candidates;

URL	description
`http://www.adria-bol.hr/`	Website of a tourist agency based in the city of Bol
`http://www.animafest.hr/`	Portal of the World Festival of Animated Film in Zagreb
`http://bol.hr/`	Tourism portal of the city of Bol
`http://www.burin-korcula.hr/`	Website of Burin, a private tourist agency Korula island
`http://www.camping.hr/`	Website of the Croatian Camping Union (CCU)
`http://www.dalmatia.hr/`	Official tourism portal of Dalmatia Country
`http://dubrovnik-festival.hr/`	Website of the Dubrovnik Summer Festival
`http://www.events.hr/`	Croatian online travel agent
`http://www.galileo.hr/`	Croatian online travel agent
`http://hhi.hr/`	Hydrographic Institute of the Republic of Croatia
`http://www.istra.hr/`	Official tourism portal of Istria
`http://www.kvarner.hr/`	Official tourism portal of Kvarner County
`http://plavalaguna.hr`	Website of the hotel company *Laguna Porec*
`http://www.liburnia.hr/`	Website of the hotel company *Liburnia Riviera Hotels*
`http://m.pulainfo.hr/`	Tourism portal of the city of Pula
`http://www.portauthority.hr/`	Website of the Croatian Association of Port Autorities
`http://www.putomania.com.hr`	Portal about travelling around the world
`http://www.tzg-rab.hr/`	Tourism portal about Rab island
`http://tzgrovinj.hr/`	Official tourism portal of Rovinj-Rovigno
`http://www.uniline.hr/`	Festival of urban culture
`http://urbanfestival.blok.hr/`	On-line reservation of accommodation in Croatia

Table 1: List of processed websites including the URL and a short description

- *10-best-filtered*: The same than *10-best*, but those pairs of documents with a segment-alignment score (provided by hunalign) under 0.3 are discarded;
- *1-best-filtered*: The same than *1-best*, but those pairs of documents with a segment-alignment score under 0.3 are discarded.

For these settings, we computed the success ratio obtained for identifying parallel documents by manually verifying a sample of the document pairs obtained. In addition to this quality evaluation, we wanted to obtain a quantitative measure of the amount of data crawled by each setting. However, using only the amount of parallel documents detected to this end presents a problem: bitextor and ILSP-FC adopt different strategies for discarding duplicates. While ILSP-FC discards (near-)duplicate documents, bitextor only discards documents containing exactly the same text. As a result, bitextor retrieves much more document pairs than ILSP-FC, but the degree of redundancy is much higher. In order to perform a fair comparison between both tools, we decided to measure the number of unique aligned segments and, therefore, to reduce the impact of redundancy in the data obtained by bitextor. To perform the alignment of the document pairs at the segment level, both corpora were further segmented into sentences[10] and tokenised using the scripts[11] and included in the Moses statistical ma-

chine translation toolkit (Koehn et al., 2007). Then, the tool hunalign was used for aligning the segments. Finally, segment pairs with a score lower than 0 were discarded.

3. Results and discussion

The pairs of documents detected by each setting were merged in a pool containing 10,662 unique pairs of documents. As expected, we observed a high degree of overlapping between the settings of the same tool.[12] However, only 8.5% of the document pairs in *all* were also in *10best*. This divergence is due to the different methods used by each tool to crawl the websites and to detect parallel documents, and suggests that they could be combined to obtain a bigger corpus. Table 2 shows the total amount of document pairs obtained with each setting, as well as the number of unique segments contained in these documents both in English and Croatian. The last column of the table contains the number of unique segment pairs obtained after aligning the collection of document pairs obtained with the tool hunalign.[13] It is worth noting that the relative difference between the numbers of parallel documents obtained by each setting is much higher than the relative difference between the numbers of unique aligned segment pairs. This confirms the idea that the number of document pairs is not an appropriate metric to check the amount of data obtained with each tool, as mentioned in Section 2.

From the pool of document pairs, a sample of 1,129 (about 10%) was randomly picked and checked, obtaining a total

---

[10]Both Bitextor and FC split the text in a document by using the HTML tags in it. However, it is possible to have pieces of text longer than a segment, so a second segmentation process is required.

[11]`https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl` and `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl`

[12]As already mentioned, settings *10best-filtered*, *1best-filtered*, and *reliable* are sub-sets of *10best*, *1best*, and *all*, respectively; in addition, 97.9% of the pairs of documents in *1best* also appeared in *10best*.

[13]All the data provided in Table 2 regarding segments was lowercased before removing duplicates in order to minimise the redundancy.

tool	setting	aligned documents	unique segments		unique aligned segment pairs
			English	**Croatian**	
focused crawler	all	3,294	46,226	47,370	40,431
	reliable	2,406	37,986	38,772	32,544
bitextor	10best	7,787	54,859	46,794	50,338
	10best-filtered	5,056	49,406	43,972	46,242
	1best	4,232	41,318	40,703	37,727
	1best-filtered	3,758	40,078	39,542	36,834

Table 2: Amount of document pairs obtained with each of the two settings of ILSP-FC, and for the four settings of bitextor. The table also reports the number of unique lowercased segments from the aligned documents both in English and in Croatian, and the number of unique lowercased aligned segment pairs obtained after aligning all these documents.

of 831 pairs confirmed as parallel documents by the human evaluators. Table 3 shows the success rates obtained by each setting when identifying parallel documents. These results confirm that, as expected, the *reliable* setting provides better precision than *all* for ILSP-FC, while the settings *1best* and *1best-filtered* are the most successful for bitextor. In a general comparison, *1best-filtered* overcomes all the other settings in terms of success rate. Another interesting detail is that the fraction of parallel documents in the whole sample is 73.6%, which is lower than the success rate obtained by each setting. This is due to the fact that the intersection of the pairs of documents obtained by all settings contains more parallel documents than non-parallel documents. In order to examine the intersection of each setting against the others and check the contribution of each setting to the resulting corpus, a similarity measurement was performed between the sub-corpora obtained with each setting. Thus, Table 4 shows the Jaccard index (Chakrabarti, 2003, Chapter 3) between the collections of aligned segment pairs obtained with each setting. Additionally, the last column of this table reports the Jaccard index between the corpus obtained with each setting and the resulting corpus, this is, the part of this corpus covered by each setting. These results show that the pair detectors integrated in these two tools could be considered complementary. For instance, the accuracy rates of the *reliable* setting of ILSP-FC and the *1best-filtered* of bitextor are 90.76% and 94.79% respectively while only 13,44% of the delivered unique segment pairs are common. Hence, it seems logical to use both tools in parallel to maximise the amount of parallel data collected from a collection of websites. Comparing the results regarding the Jaccard index of each setting with the whole corpus obtained, we can conclude that the contribution of both ILSP-FC and bitextor is quite balanced.

4. Error analysis

We devoted some time to check which were the main errors made by each tool when detecting parallel documents and some patterns were observed. Typical errors were:

- *content similarity*: Some of the websites crawled were prone to contain very similar web pages. For example, in the case of hotel chains, it is usual to find web pages about different hotels, where most of the text is

tool	setting	success rate
focused crawler	all	73.86%
	reliable	90.76%
bitextor	10best	74.70%
	10best-filtered	83.57%
	1best	92.68%
	1best-filtered	94.79%

Table 3: Results on the manual revision of detected parallel documents. For each setting, number of pairs of documents detected which were confirmed o be parallel.

the same and only a few data (name, address, number of rooms, etc.) changes. These similarities in the content caused many wrong document alignments, which were more usual in the case of bitextor, which does not remove near-duplicate documents. It is worth noting that these errors at the level of document alignment are not so severe when the corpus is aligned at segment level, since most of the aligned segment pairs are correct.

- *URL similarity*: In the case of ILSP-FC, websites keeping a highly similar URL structure caused also wrong alignments, since one of the strategies adopted by this tool is to compare URLs ignoring the differences in the content of the pages.

5. Resulting corpus

Two parallel English–Croatian corpora were obtained as a result of this work: a general corpus resulting from the union of all the 10,662 pairs of documents obtained by each setting, and a human-verified corpus resulting of the compilation of all the 831 documents confirmed as parallel by the human evaluators. These corpora are available at `http://redmine.abumatran.eu/projects/en-hr-tourism-corpus` aligned at the segment level[14] and formatted following the TMX stan-

---

[14]The alignment was performed following the methodology described in Section 2.

		Jaccard index between aligned corpora						
		focused		*bitextor*				*merged*
		all	reliable	10best	10best-filtered	1best	1best-filtered	
focused	all	—	70.84%	10.93%	11.38%	12.04%	12.06%	46.46%
crawler	reliable	—	—	11.69%	12.28%	13.19%	13.22%	37.40%
bitextor	10best	—	—	—	86.62%	68.28%	67.12%	57.84%
	10best-filtered	—	—	—	—	72.23%	73.40%	53.14%
	1best	—	—	—	—	—	95.34%	43.35%
	1best-filtered	—	—	—	—	—	—	42.33%

Table 4: Jaccard index measuring the similarity between the different collections of unique segment pairs obtained with each setting. The final column measures the Jaccard index of each setting with the *merged* corpus obtained when producing the union of all the settings.

dard.[15] In addition, a field *prop*[16] was added to each unit in the TMX file containing a comma-separated list with the names of the settings which produced it. This information is aimed at allowing to extract customised sub-corpora with different degrees of quality, depending on the settings included. After alignment, we obtained 87,024 aligned segments for the general corpus, and 9,387 for the human-verified corpus.

6. Concluding remarks

In this work we compared two tools for automatically crawling parallel corpora from multilingual websites: Focused Crawler and Bitextor. We used both tools for crawling 21 websites in the tourism domain in order to build an English–Croatian domain-specific corpus. We used several settings for crawling with each tool in order to compare them in terms of amount of parallel data obtained and precision in parallel document crawling. Our experiments proved that both tools can obtain similar precision and amount of data depending on the setting chosen. In addition, we proved that both tools obtain parallel data from different parts of the websites and, therefore, combining the corpora obtained by them allows us to mine parallel documents more exhaustively.

We finally obtained a parallel corpus consisting of 10,662 pairs of documents, which, after segment alignment, resulted in a collection of 87,024 unique pairs of segments. In addition, the human verification performed for evaluating precision allowed us to produce a smaller high-quality parallel corpus consisting of 831 pairs of documents, which were manually verified as parallel documents. After aligning this second corpus at the level of segments, we obtained 9,387 unique pairs of segments.

7. Acknowledgements

---

[15] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html
[16] http://www.gala-global.org/oscarStandards/tmx/tmx14b.html#prop

8. References

Antonova, Alexandra and Misyurev, Alexey. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144, Portland, Oregon. Association for Computational Linguistics.

Barbosa, Luciano, Rangarajan Sridhar, Vivek Kumar, Yarmohammadi, Mahsa, and Bangalore, Srinivas. (2012). Harvesting parallel text in multiple languages with limited supervision. In *Proceedings of COLING 2012*, pages 201–214, Mumbai, India.

Bojar, Ondřej, Buck, Christian, Callison-Burch, Chris, Federmann, Christian, Haddow, Barry, Koehn, Philipp, Monz, Christof, Post, Matt, Soricut, Radu, and Specia, Lucia. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

Chakrabarti, Soumen. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann.

Chen, Jisong, Chau, Rowena, and Yeh, Chung-Hsing. (2004). Discovering parallel text from the world wide web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, volume 32 of *ACSW Frontiers'04*, pages 157–161, Dunedin, New Zealand.

Désilets, Alain, Farley, Benoit, Stojanovic, M, and Patenaude, G. (2008). WeBiText: Building large heterogeneous translation memories from parallel web content. In *Proceedings of Translating and the Computer*, pages 27–28, London, UK.

Diab, Mona and Resnik, Philip. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 255–262, Philadelphia, Pennsylvania.

Esplà-Gomis, Miquel and Forcada, Mikel L. (2010). Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–

86.

Harris, Brian. (1988). Bi-text, a new concept in translation theory. *Language Monthly*, 54:8–10.

Hong, Gumwon, Li, Chi-Ho, Zhou, Ming, and Rim, Hae-Chang. (2010). An empirical study on web mining of parallel data. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 474–482, Beijing, China. Association for Computational Linguistics.

Jiang, Long, Yang, Shiquan, Zhou, Ming, Liu, Xiaohua, and Zhu, Qingsheng. (2009). Mining bilingual data from the web with adaptively learnt patterns. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 2 of *ACL'09*, pages 870–878, Suntec, Singapore.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondřej, Constantin, Alexandra, and Herbst, Evan. (2007). Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL'07, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the X Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Koehn, Philipp. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Kohlschütter, Christian, Fankhauser, Peter, and Nejdl, Wolfgang. (2010). Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450, New York, NY, USA.

Ljubešić, Nikola and Erjavec, Tomaž. (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In Habernal, Ivan and Matousek, Václav, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer.

Lui, Marco and Baldwin, Timothy. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL'12, pages 25–30, Jeju Island, Korea.

Ma, Xiaoyi and Liberman, Mark. (1999). Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, pages 538–542, Singapore, Singapore.

Mehdad, Yashar, Negri, Matteo, and Federico, Marcello. (2011). Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1 of *HLT'11*, pages 1336–1345, Portland, Oregon. Association for Computational Linguistics.

Melamed, Dan I. (2001). *Empirical methods for exploiting parallel texts*. MIT Press.

Nie, Jian-Yun, Simard, Michel, Isabelle, Pierre, and Durand, Richard. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'99, pages 74–81, Berkeley, California, USA. ACM.

Papavassiliou, Vassilis, Prokopidis, Prokopis, and Thurmair, Gregor. (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. Association for Computational Linguistics.

Resnik, Philip and Smith, Noah A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Roukos, Salim, Graff, David, and Melamed, Dan. (1995). Hansard French/English. Linguistic Data Consortium. Philadelphia, USA.

San Vicente, Iaki and Manterola, Iker. (2012). PaCo2: A fully automated tool for gathering parallel corpora from the web. In Chair), Nicoletta Calzolari (Conference, Choukri, Khalid, Declerck, Thierry, Doan, Mehmet Uur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, LREC'1, Istanbul, Turkey. European Language Resources Association (ELRA).

Shi, Lei, Niu, Cheng, Zhou, Ming, and Gao, Jianfeng. (2006). A DOM tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL'06, pages 489–496, Sydney, Australia.

Sin, Chunyu, Liu, Xiaoyue, Sin, KingKui, and Webster, Jonathan J. (2005). Harvesting the bitexts of the laws of Hong Kong from the web. In *Proceedings of the Fifth Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network*, pages 71–78, Jeju, South Corea.

Sridhar, Vivek Kumar Rangarajan, Barbosa, Luciano, and Bangalore, Srinivas. (2011). A scalable approach to building a parallel corpus from the web. In *Interspeech*, pages 2113–2116, Florence, Italy.

Tiedemann, Jörg. (2009). News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Utiyama, Masao, Kawahara, Daisuke, Yasuda, Keiji, and Sumita, Eiichiro. (2009). Mining parallel texts from mixed-language web pages. In *Proceedings of the XII Machine Translation Summit*, Ottawa, Ontario, Canada.

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai,

András, Trón, Viktor, and Nagy, Viktor. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Yan, Zhenxiang, Feng, Yanhui, Hong, Yu, and Yao, Jianmin. (2009). Parallel sentences mining from the web. *Journal of Computational Information Systems*, 6:1633–1641.

Zhang, Ying, Wu, Ke, Gao, Jianfeng, and Vines, Phil. (2006). Automatic acquisition of Chinese–English parallel corpus from the web. In Lalmas, Mounia, MacFarlane, Andy, Rger, Stefan, Tombros, Anastasios, Tsikrika, Theodora, and Yavlinsky, Alexei, editors, *Advances in Information Retrieval*, volume 3936 of *Lecture Notes in Computer Science*, pages 420–431. Springer Berlin Heidelberg.