AUTOMATIC ACQUISITION OF BILINGUAL LANGUAGE RESOURCES

Nikos Mastropavlos
Institute for Language and
Speech Processing/Athena RIC
nmastr@ilsp.gr

Vassilis Papavassiliou
Institute for Language and
Speech Processing/Athena RIC
vpapa@ilsp.gr

ABSTRACT

This paper discusses methods for automatic acquisition of bilingual corpora from the Web. Given the vast number of documents available online, the Web could be considered an excellent pool for extraction of valuable data for linguistic purposes. Therefore, methods for creating such corpora, especially when targeting less-resourced languages like Greek, can be of great value. Besides presenting a general workflow for constructing collections from the Web, this article describes our work to produce collections of English/Greek comparable documents in the "Political News", "Technological News", "Sport News", and "Renewable Energy" domains and parallel resources in the "Environment" and "Labour Legislation" domains.

1. Introduction

There is a growing literature on using the Web for constructing large comparable and parallel collections. Such resources can be used by linguists studying language use and change (Kilgarriff and Grefenstette, 2003), and at the same time be exploited in applied research fields like machine translation, cross-lingual information retrieval, multilingual information extraction, etc. Moreover, these large collections of raw data can be automatically annotated and used to produce, by means of induction tools, a second order or synthesized derivatives: rich lexica (with morphological, syntactic and lexico-semantic information) and massive bilingual dictionaries (word and multiword based) and transfer grammars.

We adopted two different strategies in order to acquire large-scale collections from which many parallel sentences could be extracted. The first method, used for the "Political News", "Technological News", "Sport News", and "Renewable Energy" domains, employs a focused crawler, i.e. an engine which starts from a few seed URLs and "travels" on the Web to find web pages in the targeted languages and relevant to specific domains (Menczer et al, 2004).

The second approach, used for acquiring parallel resources in the "Environment" and "Labour Legislation" domains, exploits a similar crawler that targets known bilingual web sites and extracts pairs of documents that are likely translations of each other. Both approaches integrate tools for text normalization, language identification, and text classification. Following the crawling process, our system filters stored pages in order to discard duplicates and keep only documents rich in textual information. The implemented components for corpus acquisition are available as web services at http://nlp.ilsp.gr/soaplab2-axis/ .

The organization of the rest of the paper is as follows: In Section 2, we refer to recent related work. In Section 3, we describe in detail the proposed workflow to construct comparable corpora. A modification of this workflow with the purpose of acquiring parallel data is presented in Section 4. Conclusions and future work are discussed in Section 5.

2. Related Work

Several publications concerning the exploitation of comparable and parallel corpora can be found in the related literature. A framework for exploiting comparable and parallel corpora for generating named entity translation pairs was introduced by Hassan et al. (2007). Tao and Zhai (2005) presented a method for examining frequency correlations of words in different languages in comparable corpora in order to find mappings between words and achieve cross-lingual information integration. Munteanu (2006) attempted to extract parallel sub-sentential fragments from comparable bilingual corpora using a signal-processing approach for producing training data sets for MT systems.

On the other hand, the number of available publications that address the issue of building such corpora is very limited. A report on different methodologies used to collect small-scale corpora in nine language pairs and various comparability levels was reported by Skadiņa et al. (2010) and the collected corpora were investigated for defining criteria and metrics of comparability. Resnik and Smith (2003) considered the Web as a parallel corpus and proposed a method based on the similarities of the HTML source in order to detectparallel web pages. A similar approach was implemented recently by Esplà-Gomis and Forcada (2010) and was delivered as the open source project called Bitextor[1].

Early approaches were based on readily available resources. Sheridan and Ballerini (1996) introduced an approach for multilingual information retrieval, applying thesaurus-based query-expansion techniques over a collection of documents provided by the Swiss news agency. Braschler and Scäuble (1998) presented a corpus-based approach for building comparable corpora using the TREC CLIR data while Talvensaari et al. (2007) presented a study which described how a comparable corpus was built from articles by a Swedish news agency and a U.S. newspaper. An initial work on acquiring comparable corpora from the web was reported by Utsuro et al. (2002). They collected articles in Japanese and English from News web sites and attempted to align them based on their publication dates. Ion et al. (2010) presented a customizable application that could be used for building comparable corpora from Wikipedia and the web by merging and organizing different web crawlers. Talvensaari et al. (2008) used a focused crawling system to produce comparable corpora in the genomics domain in English, Spanish and German languages. Even though our work follows the same methodological approach, two critical differences are: i) less-resourced languages are targeted, which significantly increases the challenge of this task and ii) a number of different topical domains are crawled extensively to produce the final results.

3. Building Comparable Corpora

Based on the assumption that documents in different languages and in a specific narrow domain could be considered comparable, we adopted a strategy consisting of two separate monolingual domain-specific crawls, one for each language. Given that each crawl provides documents relevant to the narrow domain, we believe that several parallel sentences or phrases could be extracted from the final bilingual collection. The workflow for acquiring monolingual domain-specific data is illustrated in figure 1 and each module and required resource of the focused monilingual crawler (FMC) is discussed in the following subsections. FMC is available as web service at http://nlp.ilsp.gr/soaplab2-axis/ under the name *ilsp_fmc*.

3.1 Construction of Topic Definitions

A critical issue in focused web crawling is the creation of the topic definition, since each web page visited by the crawler should be classified as relevant to the topic or not with respect to this definition. To this end, we adopted a strategy followed by many researchers (Ardö and Golub, 2007; Dorado, 2008), i.e. to use triplets (<term, relevance weight, topic-class>) as the basic entities of the topic definition. The relevance weight is a manually given score, positive or negative; a positive score indicates the power of a given term to more or less closely related to the topic; a negative score denotes terms that are likely to occur in documents of similar but not relevant domains. For example, the term “biodiversity” is closely related to the “Environment” domain and so has a large positive weight, while the term “heavy metal” could have a low positive weight denoting that this term is a common term in documents about deterioration of the environment, or a negative weight since this term is frequently met in documents about music genres. Topic-classes correspond to possible sub-categories of the target domain. For instance, “labour law and

[1] http://bitextor.sourceforge.net/

labour relations" and "personnel management and staff remuneration" could be two sub-classes of the "Labour Legislation" domain. By using the topic-classes each document under consideration is not only classified as relevant to the domain or not, but it is further categorized into a specific sub class. Introducing sub classes in a topic definition effectively prevents the bias of the collection to a specific sub class.

Topic definitions can be constructed by manually selecting a representative set of words or multiword expressions. Online resources (e.g. Eurovoc multilingual thesaurus[2] was employed in our work) provide sets of terms in different languages assigned in specific thematic categories and therefore can greatly assist in this process. Alternatively, a topic definition can be automatically extracted by small topic-specific corpora using tf-idf and term extraction algorithms.

3.2 Construction of Lists of Seed URLs

Similarly, an initial seed URL list for the English language was assembled during the topic definition construction. To expand this list, we employed a custom version of the BootCat toolkit (Baroni and Bernardini, 2004). Using random tuples (i.e. n-combinations of terms) from the terms included in the topic definition, queries were run on the Google search engine and the resulting URLs were added to the seed list. Finally, for the other languages, native speakers were asked to manually select URLs that point to relevant web pages.

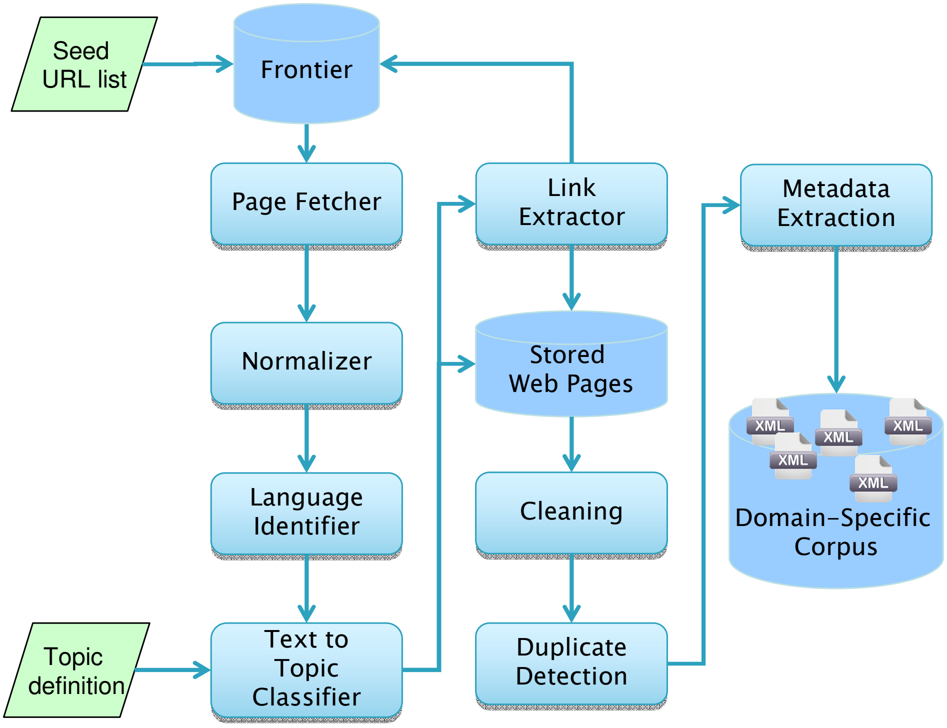


Figure 1. Workflow for acquiring monolingual domain-specific data

[2] http://eurovoc.europa.eu/

3.3 Focused Monolingual Crawling

The focused monolingual crawling is an iterative procedure that includes several steps of processing the content of the visited web pages. The schedule of the FMC is called the Frontier; the list of URLs to be fetched in each iteration. Initially, the Frontier contains the seed URLs. Then, the FMC visits these web pages, extracts their links and estimates a relevance score for each of these links (see subsection 3.3.3 below). Next, the links are sorted according to these scores and the most promising links (i.e. the links originated from very relevant pages, or contain terms in their surrounding texts) are selected and fill the Frontier for the next iteration.

3.3.1 Text Normalization

The text normalization phase involves detection of the formats and text encodings of the downloaded web pages as well as conversion of these pages into a unified format (plain text) and text encoding (UTF-8).

3.3.2 Language Identification

In the language identification phase, each downloaded web page is analysed and its language is identified. Documents that are not in the target language are then discarded. Lingua:Identify , an open-source and flexible language identifier based on n-grams, is used for this task. Lingua:Identify did not originally support the Greek language; we provided the author of the tool with a small corpus of Greek texts (taken from JRC Acquis) and a new version of the identifier was released and used throughout the subsequent work. In order to remove parts of text that are not in the targeted language, the embedded language identifier applied at paragraph level as well. Any paragraph that is not in the targeted language is annotated paragraph as “out of interest”.

3.3.3 Text Classification

The next process in the proposed pipeline is a text–to-topic classification module. Each crawled, normalized and in-target language web page is compared with the topic definition by exploiting a simple string-matching algorithm. By adopting the method described in (Ardö and Golub, 2007), the score of relevance *s* for each web page is calculated as follows:

$$s = \sum_{j=1}^{4} \sum_{i=1}^{N} \frac{w_j^l \cdot w_i^t \cdot n_{ij}}{l_j}$$

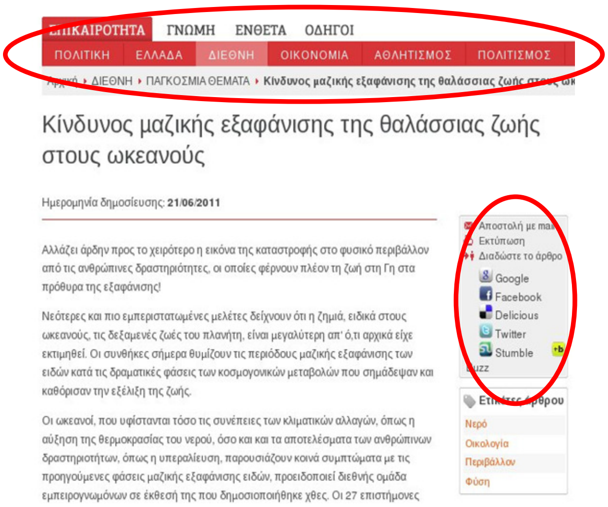
where N is the number of terms in the topic definition, w_j^l denotes the weight assigned to each location j of the HTML page (i.e. 10 for *title*, 4 for *metadata*, 2 for *keywords* and 1 for *main text*), w_i^t is the weight of term i, n_{ij} denotes the number of occurrences of term i in the location j, and l_j is the number of words in the location j.

The calculated score models the likelihood that the page under consideration contains text relevant to the target domain. Therefore, if the score of relevance is under a predefined threshold, the page is classified as irrelevant and discarded. Otherwise, the page is stored and its links are extracted and added to the list of scheduled to be visited links. The selection of a high threshold, although it ensures that the acquired pages contain rich in-domain content, have proven to sometimes “choke” the crawler since it is often the case that pages which meet this criterion cannot be found without first visiting less relevant web pages. To overcome this shortcoming, we adopted the “tunneling” algorithm according to which the crawler will not give up probing a direction immediately after it encounters an irrelevant page but will continue searching in that direction for a pre-defined number of steps. This allows the focused crawler to travel from one relevant web cluster to another when the gap (number of irrelevant pages) between them is within a limit.

Since it is not feasible to find all the web pages that exist in the Web and are relevant to a specific domain, the recall of the embedded text to topic classifier is not a critical measure. On the contrary, precision is of singificant value, since the constructed corpus consists of documents that have categorized as relevant. In order to favor precision we made the classifier stricter by introducing an additional relevance score which is based on the amount of unique terms that exist in the main content of the page.

3.3.4 Boilerplate Removal

Web pages often need to be cleaned from elements that are irrelevant to the content (see figure 2) like navigation links, advertisements, disclaimers, etc. (often called boilerplate). Since we aim to collect comparable corpora useful for linguistic purposes, such parts of the HTML source are usually redundant. Therefore, they were detected and marked as "boilerplate" by employing the Boilerpipe tool (Kohlschütter et al., 2010), which uses a set of shallow text features (link density, number of words in text blocks, etc.) for classifying individual text elements in a web page as boilerplate.



ΕΠΙΚΑΙΡΟΤΗΤΑ ΓΝΩΜΗ ΕΝΘΕΤΑ ΟΔΗΓΟΙ

ΠΟΛΙΤΙΚΗ ΕΛΛΑΔΑ ΔΙΕΘΝΗ ΟΙΚΟΝΟΜΙΑ ΑΘΛΗΤΙΣΜΟΣ ΠΟΛΙΤΙΣΜΟΣ

Αρχική › ΔΙΕΘΝΗ › ΠΑΓΚΟΣΜΙΑ ΘΕΜΑΤΑ › **Κίνδυνος μαζικής εξαφάνισης της θαλάσσιας ζωής στους ωκ**

Κίνδυνος μαζικής εξαφάνισης της θαλάσσιας ζωής στους ωκεανούς

Ημερομηνία δημοσίευσης: **21/06/2011**

Αλλάζει άρδην προς το χειρότερο η εικόνα της καταστροφής στο φυσικό περιβάλλον από τις ανθρώπινες δραστηριότητες, οι οποίες φέρνουν πλέον τη ζωή στη Γη στα πρόθυρα της εξαφάνισης!

Νεότερες και πιο εμπεριστατωμένες μελέτες δείχνουν ότι η ζημιά, ειδικά στους ωκεανούς, τις δεξαμενές ζωές του πλανήτη, είναι μεγαλύτερη απ' ό,τι αρχικά είχε εκτιμηθεί. Οι συνθήκες σήμερα θυμίζουν τις περιόδους μαζικής εξαφάνισης των ειδών κατά τις δραματικές φάσεις των κοσμογονικών μεταβολών που σημάδεψαν και καθόρισαν την εξέλιξη της ζωής.

Οι ωκεανοί, που υφίστανται τόσο τις συνέπειες των κλιματικών αλλαγών, όπως η αύξηση της θερμοκρασίας του νερού, όσο και και τα αποτελέσματα των ανθρώπινων δραστηριοτήτων, όπως η υπεραλίευση, παρουσιάζουν κοινά συμπτώματα με τις προηγούμενες φάσεις μαζικής εξαφάνισης ειδών, προειδοποιεί διεθνής ομάδα εμπειρογνωμόνων σε έκθεσή της που δημοσιοποιήθηκε χθες. Οι 27 επιστήμονες

Figure 2. Boilerplate of a web page

3.3.5 Duplicate Detection

In (near) duplicate detection each new candidate document is checked against all other documents appearing in the corpus (i.e. by document similarity measures) before being added to the collection. An efficient algorithm for deduplication, which is implemented as an open source tool, is SpotSigs (Theobald et al., 2008). The algorithm represents each document as a set of spot signatures. A spot signature is a chain of words that follow frequent words as these are attested in a corpus. SpotSigs classifies documents with respect to the cardinality of their set of spot signatures and so significantly reduces the time complexity.

3.3.6 Metadata Extraction

In this task, the HTML source of the stored web pages is scanned and the available metadata (i.e. original URL, keywords, title, etc) are extracted. In addition, the structure of the web page is detected and a special attribute (i.e. title, listitem or heading) is added to each paragraph. Parts of such an XML file are presented in the following example:

```
<?xml version='1.0' encoding='UTF-8'?>
<cesDoc version="0.4" xmlns="http://www.xces.org/schema/2003" xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <cesHeader version="0.4">
        <fileDesc>
            <titleStmt>
                <title>Danube Delta - UNESCO World Heritage Centre</title>
                <respStmt>
                    <resp>
                        <type>Crawling and normalization</type>
                        <name>ILSP</name>
                    </resp>
                </respStmt>
            </titleStmt>
            <sourceDesc>
                <biblStruct>
                    <monogr>
                        <title>Danube Delta - UNESCO World Heritage Centre</title>
                        <imprint>
                            <format>text/html</format>
                            <eAddress>http://whc.unesco.org/en/list/588</eAddress>
                        </imprint>
                    </monogr>
                </biblStruct>
            </sourceDesc>
        </fileDesc>
        <profileDesc>
            <langUsage>
                <language iso639="en"/>
            </langUsage>
            <textClass>
                <keywords>
                    <keyTerm>UNESCO</keyTerm>
                    <keyTerm>World Heritage Centre</keyTerm>
                </keywords>
                <domain></domain>
                <subdomain>natural environment</subdomain>
                <subject/>
            </textClass>
            <annotations>
                <annotation>http://sifnos.ilsp.gr/ data/201109/ENV_EN/11213.html</annotation>
            </annotations>
        </profileDesc>
    </cesHeader>
    <text>
        <body>
            <p id="p1" crawlinfo="boilerplate">jump to the content</p>
            <p id="p2" crawlinfo="boilerplate" type="listitem">English</p>
            <p id="p3" crawlinfo="boilerplate" type="listitem">Français</p>
            ...
            <p id="p61" topic="delta;marsh">The waters of the Danube, which flow into the Black Sea, form the
largest and best preserved of Europe's deltas. The Danube delta hosts over 300 species of birds as well as 45 freshwater
fish species in its numerous lakes and marshes.</p>
            <p id="p62" crawlinfo="ooi-length">Delta du Danube</p>
            <p id="p63" crawlinfo="ooi-lang">Les eaux du Danube se jettent dans la mer Noire en formant le plus
vaste et le mieux préservé des deltas européens. Ses innombrables lacs et marais abritent plus de 300 espèces d'oiseaux
ainsi que 45 espèces de poissons d'eau douce.</p>
            ...
            <p id="p236" crawlinfo="boilerplate">Not a member yet?</p>
        </body>
    </text>
</cesDoc>
```

3.4 Acquired Comparable Corpora

We employed the described workflow for automatic collection of comparable corpora. Texts were collected in the domains of Renewable Energy (RE), Political News (PN), Sport News (SN) and Technological News (TN) and in two languages: English (EN) and Greek (EL). To strengthen the comparability rank of the collected corpora, further sub-categorization was enforced on all domains. Table 1 shows the subclasses selected for each domain. Table 2 illustrates the quantities of the acquired resources in the selected language pairs and domains.

Domain	Subclasses
RE	Wind power Hydropower Solar energy Biomass Biofuel Geothermal energy
PN	Ireland bailout Google faces competition inquiry Leaving the euro
TN	iPad models, presentations and reviews iPhone models, presentations and reviews Facebook and privacy issues
SN	Football and the Barcelona – Real Madrid match (29/11/2010) Tennis and the ATP World Tour Final between Federer and Nadal Ice hockey and the Vancouver 2010 Winter Olympics tournament

Table 1 Domains and subclasses

Language pairs (L1-L2)	Domain	# of tokens in L1 (Mt)	# of tokens in L2 (Mt)
EN-EL	RE	19	0.9
EN-EL	PN	25	25.8
EN-EL	TN	25.7	10.2
EN-EL	SN	8.8	13.5

Table 2 Quantitative information for acquired comparable data. (figures in Mega tokens (Mt)).

4. Building Parallel Corpora

This section describes the required modifications of the proposed workflow in order to acquire parallel documents from the web. To this end, anothter component called Focused Bilingual Crawler (FBC) was implemented and is available as web service at http://nlp.ilsp.gr/soaplab2-axis/ under the name *ilsp_bilingual_crawl*. The first requirement concerns the seed URL list; these lists should consist of links pointing to web pages that are relevant to a predefined domain and originated from multilingual web domains. In general, the construction of such lists is a time-consuming process, which is being performed manually. Then, the crawler starts from an initial URL, and in a spider-like mode finds the links within these pages pointing to pages inside the same web site, visits the new pages and so on.

Following the same processing steps, each web page is normalized and its language is identified as explained in subsections 3.3.1 and 3.3.2 respectively. If the detected language is one among the targeted languages, the web page is further processed by the text-to-topic classifier. Since the document could be in any of the two languages, the topic definition should be bilingual (i.e. the new list of terms is the union of the two corresponding monolingual topic definitions).

To this end, the collected documents are in the targeted languages, originated from the same web site and are relevant to the selected domain. In order to identify pairs of web documents that could be considered as translations of each other, an additional module is required. A well-known tool that meets this need is Bitextor (Esplà-Gomis and Forcada, 2010). For each candidate pair of documents, the relative difference in file size, the relative difference in length of plain text, the edit distance of web page fingerprints constructed on the basis of HTML tags, and the edit distance of the lists of numbers occurred in the documents are examined. If all measures are under the corresponding thresholds, the pair under consideration is considered a pair of parallel documents. Based on the "nature" of the documents and the effectiveness of Bitextor, it is very likely that the resulting pairs consist of parallel documents.

4.1 Acquired Parallel Corpora

In order to examine the effectiveness of the modified workflow in acquiring parallel data from the web, we used it for constructing parallel corpora for the English-Greek language pair, in two domains ("Environment" and "Labour Legislation" as a wide and a narrow one, respectively). Table 3 illustrates the amount of data collected by using the proposed method.

Since the identification of web domains containing pages in the targeted languages and relevant to the selected domains is a time-consuming task and was done manually, only a few web domains were found to be crawled. However, the amount of acquired data was a good start for adapting an SMT system in these domains. In fact, training a baseline SMT system with the collected in-domain data improved the translation quality as reported in (Pecina et al., 2011).

	ENV	**LAB**
# of web domains	6	4
# of document pairs	147	126
# of tokens	17,033	13,169

Table 3 Quantitative information for acquired parallel data.

5. Conclusions and Perspectives

We have presented a method for acquiring comparable and parallel corpora from the web. Each module of the proposed workflow is detailed in both cases. The components (i.e. FMC and FBC) that have been implemented for these tasks are available as web services at http://nlp.ilsp.gr/soaplab2-axis/. By employing these tools, we collected documents in three domains for the English-Greek language pair. Although one of the targeted languages is less-resourced (i.e. Greek), the amount of acquired resources might be a useful collection for extracting a sufficient number of parallel sentences to train/adapt/test an SMT system. In order to verify this assumption, we aim to examine the collected resources and calculate the number of translation equivalents that could be extracted. A next step will be the evaluation of an SMT system adapted to the selected domains by using the web-crawled data.

Future work also concerns the enhancement of each module applied on each separate task. For example, we plan to combine the text-to-topic classifier with the cleaning tool in order to identify parts of texts which are in general domain and discard them.

6. Acknowledgements

This work is supported by the PANACEA (Un. Pompeu Fabra, Istituto di Linguistica Computazionale, ILSP, Un. of Cambridge, Dublin City Univ., ELDA, Linguatec GmbH) and ACCURAT (Tilde SIA, Un. of Sheffield, Un. of Leeds, ILSP, Un. of Zagreb, DFKI, Romanian Academy Research Institute for Artificial Intelligence, Linguatec GmbH, Zemanta), 7th Framework Research Programme of the European Union, contracts 7FP-ITC-248064 and 248347.

References

Ardö, A., and Golub, K. 2007. Focused crawler software package. Technical report.

Baroni, M., and Bernardini, S. 2004. “BootCaT: Bootstrapping corpora and terms from the web”. In Proceedings of LREC 2004. 1313-1316.

Braschler, M., and Scäuble, P. 1998. “Multilingual information retrieval based on document alignment techniques”. In ECDL ’98: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, London: Springer-Verlag, pp. 183–197.

Dorado, I. G. 2008. “Focused Crawling: algorithm survey and new approaches with a manual Analysis”. Master thesis.

Esplà-Gomis, M., and Forcada, M. L. 2010. “Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor”. The Prague Bulletin of Mathemathical Lingustics, 93, pp.77–86.

Hassan, A., Fahmy, H., and Hassan, H. 2007. “Improving named entity translation by exploiting comparable and parallel corpora”, In Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing (RANLP), AMML Workshop.

Ion, R., Tufiş, D., Boroş, T., Ceauşu, A., and Ştefănescu, D. 2010. “On-Line Compilation of Comparable Corpora and their Evaluation”. In Proceedings of the 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, University of Zagreb, Dubrovnik, Croatia, pp. 29-34.

Kilgarriff, A., and Grefenstette, G. 2003. “Introduction to the special issue on the web as corpus”, Computational Linguistics - Special issue on web as corpus archive, 29 (3), pp. 333-347.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. 2010. “Boilerplate Detection using Shallow Text Features”. The Third ACM International Conference on Web Search and Data Mining.

Menczer F., Pant G.., and Srinivasan, P. 2004. “Topical Web Crawlers: Evaluating Adaptive Algorithms”, ACM Transactions on Internet Technology, 4 (4), pp. 378–419.

Munteanu, D. S. December 2006. “Exploiting Comparable Corpora”. PhD Thesis, University of Southern California. ©2007 ProQuest Information and Learning Company.

Pecina, P., Toral, A., Way, A., Prokopidis, P., Papavassiliou, V. and Giagkou, M. 2011. “Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation”. In M.L. Forcada, H. Depraetere & V. Vadeghinste (editors), Proceedings of the 15th Annual conference of the European Association for Machine Translation, pages 297-304. Leuven, Belgium.

Resnik, P., and Smith, N. A. 2003. “The Web as a Parallel Corpus”, Computational Linguistics, 29 (3), pp. 349-380.

Sheridan, P., and Ballerini, J. P. 1996. “Experiments in multilingual information retrieval using the SPIDER system”. In SIGIR ’96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York: ACM Press, pp. 58–65.

Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufis, D., and Gornostay, T. 2010. “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. LREC, Malta. pp. 6-14.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. 2007. “Creating and exploiting a comparable corpus in cross-language information retrieval”. ACM Transactions on Information Systems 25(1), 4.

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. (2008), “Focused Web crawling in the acquisition of comparable corpora”, Information Retrieval, 11 (5), pp. 427-445.

Tao, T., and Zhai, C. 2005. “Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration”. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’05), pp. 691–696.

Theobald, M., Siddharth, J., and Paepcke, A. 2008. “SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections”. In 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008).

Utsuro, T., Horiuchi, T., Chiba, Y., and Hamamoto, T. 2002. “Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites”. In AMTA ’02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, London: Springer-Verlag, pp. 165–176.