# Automatic Acquisition of Parallel Corpora from Websites with Dynamic Content

**Yulia Tsvetkov, Shuly Wintner**

Department of Computer Science
University of Haifa
31905 Haifa, Israel
yulia.tsvetkov@gmail.com, shuly@cs.haifa.ac.il

**Abstract**

Parallel corpora are indispensable resources for a variety of multilingual natural language processing tasks. This paper presents a technique for fully automatic construction of constantly growing parallel corpora. We propose a simple and effective dictionary-based algorithm to extract parallel document pairs from a large collection of articles retrieved from the Internet, potentially containing manually translated texts. This algorithm was implemented and tested on Hebrew-English parallel texts. With properly selected thresholds, precision of 100% can be obtained.

## 1. Introduction

Parallel corpora are crucial resources for NLP applications that require some sort of semantic interpretation: machine translation, automatic lexical acquisition, word sense disambiguation, etc. Collecting corpora, representing and maintaining them are non-trivial tasks. But the main challenge is to find a good source of manually translated parallel texts. An example of such a source is translated literature, but in most cases it cannot be used due to copyright restrictions or fees. Religious texts are not a subject of intellectual property, but their language is often outdated and the domain is too specific. Other examples of possible sources of parallel corpora are translated texts produced by government agencies, software and military manuals, but the language of these documents tends to be technical and domain-specific, and the size of such corpora is limited. Parliamentary proceedings, such as Europarl (Koehn, 2005) or the Canadian Hansards, are large and valuable parallel corpora, although their content is limited to legislative discourse. Unfortunately, such corpora are unavailable for Hebrew (which is the subject of our research) and many others medium-density languages (Varga et al., 2005).

Therefore, there is a natural need to search for translated materials on the Web - "a huge fabric of linguistic data often interwoven with parallel threads" (Resnik and Smith, 2003). In this paper we describe a novel content-based algorithm to extract parallel articles from a large collection of documents retrieved from the Internet, which potentially contain manually translated texts. We compiled the first Hebrew-English parallel corpus, containing articles on news, politics, sports, economics, literature, etc. We perform a daily crawl of Web sites with dynamic contents (newspaper sites), extending our corpus constantly. The average number of parallel sentences added to our corpora every month is 3625. Evaluation results show that we obtain 100% precision and 86.5% recall (threshold values were chosen to favor precision over recall, since the quality of the corpus is crucial for us while its size is just a matter of time).

Although the experiments were held for Hebrew-English, the proposed method is independent of linguistic knowledge and can be generalized to any other language pair for which a bilingual dictionary is available. If a dictionary is unavailable, a novel multilingual dictionary PANDICTIONARY (Mausam et al., 2009) can be used. It was extracted from various dictionaries over the Web and contains over 200 million pairwise translations in over 200,000 language pairs.

## 2. Related work

Most of the existing tools that harvest a parallel corpus from a collection of texts that may contain translated documents are designed as the following pipeline:

1. Detection of Web sites that are likely to have translated materials
2. Extraction of parallel texts from these sites.

STRAND (Resnik, 1998; Resnik, 1999) is an architecture for structural translation recognition. To detect bilingual Web sites, a search engine query is used to find "parents" and "siblings": Web sites containing links to translated versions of the same site. At the next stage poor candidates are filtered out by comparing the structure (HTML tags) of two pages and the lengths of the translated texts. In a later version of STRAND (Resnik and Smith, 2003), content based matching of the texts was added. Text similarity score is computed as

$$\frac{\#\textit{word-to-word translations}}{\#\textit{word-to-word translations} + \#\textit{untranslated words}}$$

To compute the number of translations, Resnik and Smith use a symmetric word-to-word translational model (Melamed, 2000), with additional complexity improvements. This technique was tested on English-French document pairs and reported as competitive to the structure-based approach of STRAND.

In BITS (Bilingual Internet Text Search) (Ma and Liberman, 1999) candidate Web sites are defined by their domain names, e.g., .de sites are considered as candidates in German. Ma and Liberman assume additionally that 10% of these sites include translations to English, and hence use

the entire domain as a set of candidates. To detect parallel documents, the system defines the content similarity for every two texts as follows:

$$sim(A, B) = \frac{\#\textit{translation token pairs}}{\#\textit{tokens in text A}}$$

Translation token pairs within a fixed window in a parallel text are detected using a translation lexicon. Additional filters are applied for document length, similarity of anchors, etc. BITS was used to collect a 63MB corpus of English-German texts.

PTMINER (Chen and Nie, 2000) follows Resnik's technique to identify candidate sites by submitting particular requests to search engines. Then, parallel pairs are detected by filename and text length comparison, language identification and sentence alignment. English-French and English-Chinese corpora were produced with this technique.

To the best of our knowledge, none of the existing techniques was applied to Hebrew. All the architectures discussed above are designed to perform an unsupervised retrieval of a static snapshot of parallel candidate sites. We believe that this method is likely to miss the most valuable translation sources. In the next section we explain this claim along with an alternative approach: to manually detect candidate sites, and then automatically monitor them over time. Moreover, we describe a novel content-based algorithm for parallel text matching and its application to the Hebrew-English language pair.

## 3. Acquisition of Parallel Corpora

### 3.1. Articles content and availability

In order to retrieve quality parallel corpora, texts should be searched on sites that are not biased to a specific subject and not edited by the same person. In addition, to guarantee the continuous growth of the corpus, sites with dynamic content should be used. Newspaper sites satisfy both conditions: they cover a wide variety of domains: politics, culture, science, sports, arts and leisure, etc.; and new articles are published frequently. Identification of such sites can be done manually, since there are few such sites and even one or two are sufficient to build a good resource. Due to the dynamic nature of these sites the size of the corpus is just a matter of time. Previously proposed techniques for automatic detection by querying search engines are unlikely to find such sites: articles usually do not contain links to their translated version, since these versions are targeted to a different readership. Translated articles can be located on different domains and maintained by different teams, and their URL does not necessarily contain the title of the article or any other identification of its identity. Therefore, neither HTML structure nor filename are useful features for article comparison, and detection of document pairs can only be done by semantic analysis of the texts.

As a source for building our corpus we use a daily on-line newspaper in Hebrew and its version in English. Not all articles are translated, and some are only translated partially.

### 3.2. Parallel Corpora Builder

Our system, Parallel Corpora Builder (PCB), was developed to collect a parallel corpus from websites with dynamic content which potentially contain translated texts. The system architecture is illustrated in Figure 1. In the following subsections we describe our system in detail.

#### 3.2.1. Web crawling

A Cron job is used to run a crawler several times a day and to harvest all fresh articles. Web crawling of the sites is a purely technical problem. We use a simple script to clean downloaded web pages from HTML tags and extract only text and metadata (date, domain, source URL, etc.)

The following features facilitate the task of collecting newspaper articles:

- To locate links to recently published articles, we use RSS feeds that are usually available on newswire sites.
- On-line newspaper articles commonly contain a link to the print version. We download these pages instead of the original articles, since they usually contain less user interface components such as Javascript, Flash, etc., and therefore require smaller effort to extract the raw text.

#### 3.2.2. Identification of parallel articles

We run a content-based comparison of all Hebrew-English document pairs that were collected during the previous month to extract translated documents. Two documents $E, H$ are defined as mutual translations, if $E$ contains enough translated terms from $H$ and vice versa. We now detail this process.

Morphological analysis tools for Hebrew (Itai and Wintner, 2008) and for English (Minnen et al., 2001) are used to reduce inflected forms of words to a common base form. Then, after tokenization, lemmatization and stop word removal, each article is represented by its bag of words (BOW). We then generate a BOW that represents the translation of this article to the parallel language. We use the same dictionary a in both directions; in our case, this is a small Hebrew-English dictionary consisting of some 20,000 handcrafted translation pairs, augmented by some 40,000 automatically extracted ones (Itai and Wintner, 2008). A translated BOW is usually much bigger than the one in the original language, since all possible translations of each word are added. Given a Hebrew-English text pair, we have

- $H$ - the BOW of the Hebrew text
- $H2E$ - the BOW of translations of $H$ to English
- $E$ - the BOW of the English text
- $E2H$ - the BOW of translations of $E$ to Hebrew

the two texts are identified as mutual translations and added to the parallel corpus if they satisfy the following formula:

$$(\frac{|H \cap E2H|}{|H|} > T_{Heb}) \textit{ and } (\frac{|E \cap H2E|}{|E|} > T_{Eng})$$

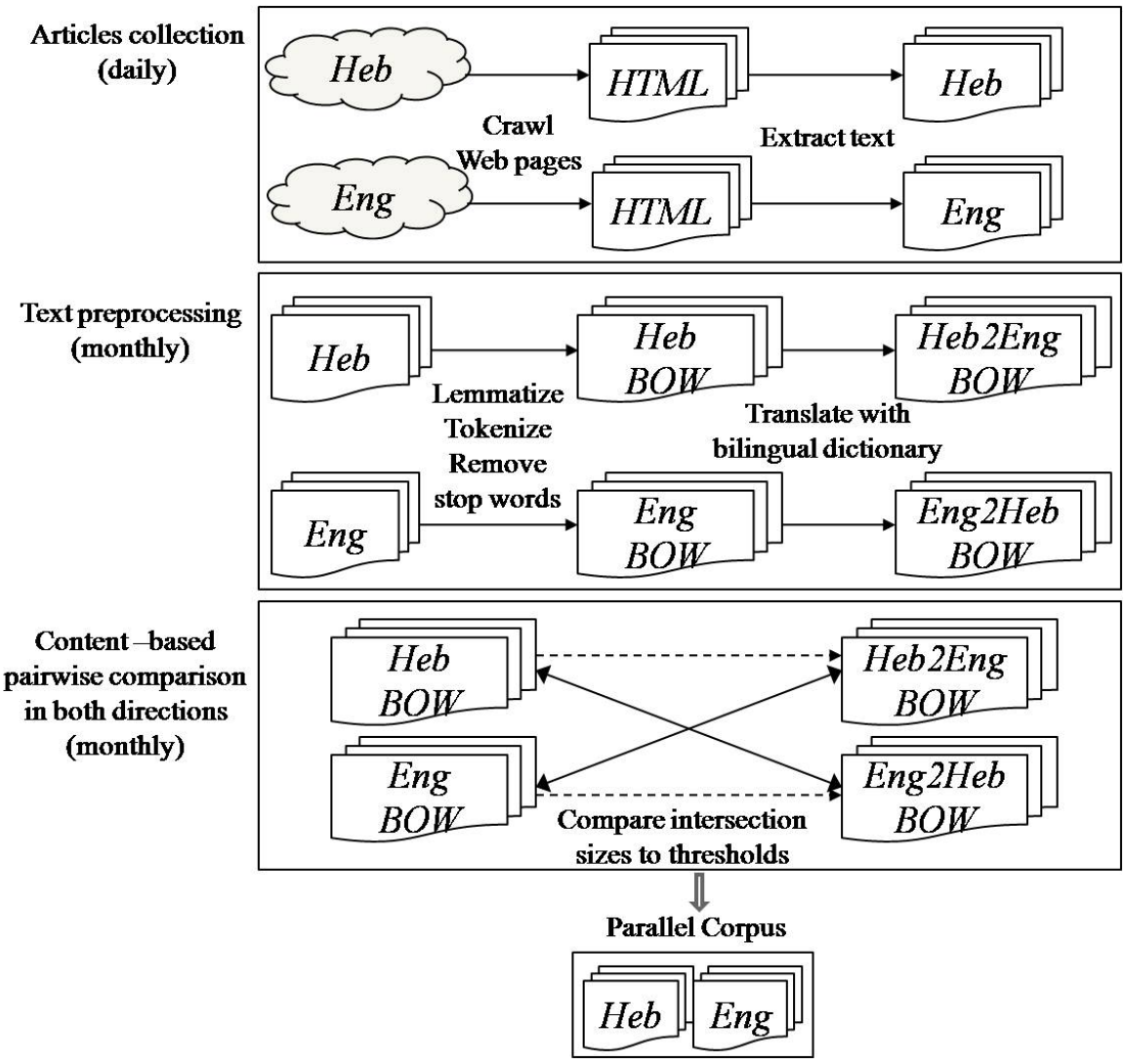


Figure 1: Parallel Corpora Builder (PCB) architecture.

where $T_{Heb}$ and $T_{Eng}$ are threshold values for Hebrew and English documents, respectively. The values of the thresholds were determined empirically, based on data collected in the first month, to maximize precision (see Section 4.).

Our experiments show that if text similarity is computed only in one direction, many false positives are added, and tuning the threshold value does not resolve this problem: for tighter thresholds, translated texts are filtered out along with the false positives. Bidirectional similarity check shows a dramatical improvement of translation detection resulting in perfect precision. In addition, the bidirectional approach is useful to filter out partially translated texts.

Moreover, to achieve perfect precision, we apply the following policy to texts that have more than one parallel document (this is a very rare case): we remove this text with all its candidate translations from the parallel corpus. The only case of the above scenario is when these articles are very closely related in subject.

Since we compare all possible pairs of documents, complexity may become a serious obstacle for large amounts of data. To solve this problem we rely on the fact that translated articles are published on the site in relatively close time intervals. We split the downloaded data to groups, stamped by the time they appeared on the Web site. Then, we run the pair detection algorithm monthly: every month we collect on average about 1500 articles in Hebrew and 600 in English, and comparison of all pairs is feasible.

## 4. Evaluation

The evaluation was performed on Hebrew and English articles collected during 3 months. As we mention above, we deliberately favor precision over recall, and our system was designed to filter out all suspicious documents. To compute the recall, we ran our system with lower thresholds and manually checked the results, to identify undetected translations. Table 1 details the evaluation results.

The main advantage of our algorithm is its simplicity: without sophisticated heuristics or probabilistic models, we use the naive BOW comparison and achieve excellent results. Indeed, the extracted corpus has been proven useful for identifying Hebrew (and English) multi-word expressions (Tsvetkov and Wintner, 2010).

| Month | English articles | Hebrew articles | Parallel articles | Detected parallel articles | Precision | Recall |
|---|---|---|---|---|---|---|
| 07 | 624 | 1530 | 168 | 145 | 100% | 86.3% |
| 08 | 548 | 1486 | 172 | 149 | 100% | 86.6% |
| 09 | 600 | 1341 | 165 | 143 | 100% | 86.7% |
| average | 573 | 1452 | 168 | 145 | 100% | 86.5% |

Table 1: PCB evaluation

## Acknowledgments

We wish to thank Gennadi Lembersky for his help. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grants No. 137/06, 1269/07).

## 5. References

Jiang Chen and Jian-Yun Nie. 2000. Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28, Morristown, NJ, USA. Association for Computational Linguistics.

Alon Itai and Shuly Wintner. 2008. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, March.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X, Phuket, Thailand*.

Xiaoyi Ma and Mark Liberman. 1999. BITS: A method for bilingual text search over the web. In *Machine Translation Summit VII, Singapore*.

Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. 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*, pages 262–270, Suntec, Singapore, August. Association for Computational Linguistics.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26:221–249.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223.

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

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In *AMTA '98: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup*, pages 72–82, London, UK. Springer-Verlag.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 527–534, Morristown, NJ, USA. Association for Computational Linguistics.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. Submitted.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596, Borovets, Bulgaria.