A Scalable Approach to Building a Parallel Corpus from the Web

Vivek Kumar Rangarajan Sridhar, Luciano Barbosa, Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932
vkumar,lbarbosa,srini@research.att.com

Abstract

Parallel text acquisition from the Web is an attractive way for augmenting statistical models (e.g., machine translation, cross-lingual document retrieval) with domain representative data. The basis for obtaining such data is a collection of pairs of bilingual Web sites or pages. In this work, we propose a crawling strategy that locates bilingual Web sites by constraining the visitation policy of the crawler to the graph neighborhood of bilingual sites on the Web. Subsequently, we use a novel recursive mining technique that recursively extracts text and links from the collection of bilingual Web sites obtained from the crawling. Our method does not suffer from the computationally prohibitive combinatorial matching typically used in previous work that uses document retrieval techniques to match a collection of bilingual webpages. We demonstrate the efficacy of our approach in the context of machine translation in the tourism and hospitality domain. The parallel text obtained using our novel crawling strategy results in a relative improvement of 21% in BLEU score (English-to-Spanish) over an out-of-domain seed translation model trained on the European parliamentary proceedings.

Index Terms: Web crawling, parallel text, machine translation

1. Introduction

Speech and text acquisition from the Web has been a topic of interest in several speech and language applications recently. The primary motivation for acquiring data from the Web has been to augment statistical models to either improve the coverage (thereby reducing out-of-vocabulary rate) or incorporate in-domain knowledge into the models. Augmenting Web text as part of language model in speech recognition has demonstrated significant performance improvements (in terms of word error rate) [1]. While techniques for monolingual text acquisition from the Web has been well established, acquiring parallel text from the Web is still a challenging problem.

Parallel texts are translations of the same text in different languages. Typically, the process of procuring parallel text from the Web comprises three steps. First, bilingual Web sites are identified either manually [2] or through a search engine [3]. Second, the pages in the bilingual Web sites are aligned using either structural cues [3] or document retrieval techniques [4, 5]. Finally, the sentences in the paired pages (documents) are aligned using dynamic programming. Some approaches omit the first step by directly trying to find pairs of webpages (documents) from a huge collection of indexed pages from the Web [5]. The drawback of previous work on acquiring parallel text from the Web is that the quality and scale of parallel data is dependent on the initial pairs of bilingual Web sites that are often difficult to obtain on a large scale. Furthermore, matching pages using document retrieval techniques can be computationally prohibitive due to the combinatorial comparisons.

Statistical machine translation has been one area where the need for such bilingual text has engendered several approaches for mining parallel text from the Web. In this paper, we focus on automatically identifying bilingual Web sites and subsequently the parallel pages within the bilingual sites. Our solution is based on the assumption that bilingual Web sites present a parallel link structure. More precisely, the Web site's link structure of pages in a particular language is similar to the link structure of pages in other language. Based on this assumption, we present a novel intra-site crawling approach that recursively identifies pairs of links (denoting parallel text) from an initial pair of bilingual Web site root links using dynamic programming. In contrast with previous work that looks for document pairs across the entire set of pages on the Web site [6], our approach reduces the search space for parallel pages significantly since its search is restricted to the outlinks of each pair of nodes in the graph representing bilingual Web sites. The recursive procedure is highly parallelizable and facilitates accelerated intra-site crawling.

Experimental results using our proposed methodology for acquiring parallel text results in significant improvements in machine translation accuracy. The improvements are demonstrated in the context of augmenting large out-of-domain MT models with in-domain Web bitext. Our procedure does not require a machine translation system in any step. We require only a word lexicon that is either available freely or obtained through automatic alignment of out-of-domain training data as seed.

2. Bilingual Web Crawler

The goal of the bilingual Web crawler is to locate bilingual sites on the Web. It is composed of three components: inter-site crawler, bilingual site detector and entry point identifier. We explain these components briefly in the following sections. Detailed information regarding the inter-site crawler and bilingual site detector can be found in [7].

2.1. Inter-Site Crawling

We implemented our strategy for locating bilingual sites by imposing the constraint that the crawler stay in the Web neighborhood graph of bilingual sites that are progressively discovered by the crawler. More specifically, the crawler explores the neighborhood graph defined by the bipartite graph composed by the backlink pages (BPs) of bilingual sites and the pages pointed by BPs (forward pages). *Backlinks* of a page p are the links that

point to p and *outlinks (forward links)* are the links that p points to. This strategy is based on the findings that Web communities are characterized by directed bipartite subgraphs [8]. Our assumption is that the Web region comprised by this bipartite graph is rich in bilingual sites as backlink pages typically point to multiple bilingual sites.

Retaining the crawler in the graph neighborhood of bilingual sites (the bipartite graph) is our first attempt towards an effective search for such sites. However, there may be many links in the graph that do not lead to relevant sites. In order to identify promising URLs in the two different page sets of the bipartite graph, we employ supervised learning. For each set (backlink and forward sets), the crawler builds a classifier that outputs the relevance of a given link in that particular set. Relevant links in the forward pages' set represent URLs of bilingual sites, i.e., links that give immediate benefit, whereas relevant links in the backlink pages' set are URLs of backlink pages that contain outlinks to bilingual sites (delayed benefit). The seeds used to initialize the crawler were obtained from Open Directory Project [1]. For our experiments, we selected a subset of 1000 random links in the sections related to Spanish speaking countries over multiple crawls.

2.2. Bilingual Site Detection

Once the inter-site crawler discovers potential bilingual sites, the next step is to verify their bilinguality. We accomplish this using a two step approach. First, we use a supervised learning to predict if a given page has links to parallel text (Link Predictor). Second, we verify whether the pages whose URLs are considered relevant by the Link Predictor are in different languages.

The role of the Link Predictor is to identify links that point to parallel text in a Web site. Similar to previous approaches, we explore patterns in the links, but instead of creating a predefined list [6], we use supervised learning to perform this task. Our assumption is that pages of bilingual sites contain some common link patterns. For instance, pages in English might have a link to its version in Spanish, containing words such as "espanol" and "castellano" in its anchor, URL, etc. In essence, the Link Predictor works as a low-cost filter, its cost is associated with the link classifications which is very low.

In the second step of the bilingual site detection, the detector verifies if the pages whose links were considered relevant by the Link Predictor are in the languages of interest. The language identification is then performed in all pages of that candidate list and if different pages are in the language of interest, the site is considered as bilingual. Since the subsequent steps assume high reliability of the initial root links, it is important to note that the bilingual site detection produces a very high-precision collection of bilingual sites (it obtained precision higher than 90%) and a low cost (only 2 to 3 pages per site were necessary to download).

2.3. Identification of entry pairs

The final step of bilingual Web crawling involves identifying pairs of pages that represent the entry points of parallel text in the bilingual Web sites. For this task, we leverage information from the link prediction and language identification (see Section 2.2). The Link Predictor identifies a set of candidate links that point to parallel texts in the Web site and then performs language identification of the pages that are referred to by these candidates. We assume that the entry points to parallel text in the Web site are contained in this set of candidate pages. Hence, the problem reduces to matching pages across source and target language that are translations of each other.

Let $P = p_1, \ldots, p_{|P|}$ be the set of $|P|$ candidate pages identified by the Link Predictor and P_s, P_t denote the pages in the source and target language, respectively ($|P| = |P_s| + |P_t|$). Our objective is to identify the closest translation of a page in P_s into a page in P_t, i.e, the entry pair $(\hat{p}_s, \hat{p}_t)$.

$$(\hat{p}_s, \hat{p}_t) = \underset{p_s \in P_s, p_t \in P_t}{\arg\max} \; (similarity(trans(p_s), p_t)) \quad (1)$$

where p_s is a page in P_s and p_t in P_t; $similarity$ is a function that calculates the similarity between two pages and $trans$ translates a given page from the source language to the target language.

In our implementation, we use a word-based approach for picking the most likely translation of a word in the target language given by a dictionary. The dictionary was obtained automatically by performing word alignment of Europarl corpus (see Section 3). Once the source page is translated, the similarity between two pages is measured using Jaccard similarity [9] and the two most similar pages are selected as the entry points. Applying the bilingual Web crawling technique generated a set of 20,186 bilingual Web sites that are potential entry points to parallel text. In the next section, we describe a recursive approach to mining the initial collection of bilingual pairs to acquire parallel text.

3. Recursive Intra-Site Crawling

One of the main drawbacks of previous work that use document matching procedure to align parallel webpages [4, 5] is the high computational cost. For example, the naive method of cross-lingual document matching is quadratic in the number of documents. Approximate matching using only top n-grams instead of the entire document [5] or heuristics such as closeness of publication dates [4] is typically used to alleviate the complexity. In contrast, our bilingual Web crawl generates a high quality set of bilingual root sites, obviating the need to perform subsequent combinatorial matching. We take advantage of the reliability in the root bilingual sites to devise a recursive procedure for extracting parallel text. Our approach is based on the assumption that the link graphs of different languages in a bilingual Web site are similar.

First, we extract the text and links contained in each pair of bilingual root Web sites. Since the text obtained from Web sites are typically in paragraph format, we used a sentence segmenter similar to [10] to segment the sentences on both the source and target side. The sentence segmenter is a maximum entropy classifier that predicts sentence boundaries by using local contextual features at each word boundary. Next, the sentences and links from the initial pair of bilingual Web sites are aligned using dynamic programming. The aligned links are then used as the next set input pairs while the aligned sentences accrue to compose the parallel corpus. The procedure is illustrated in Figure 1.

We used a lexicon-based sentence alignment that uses a dynamic programming method to find the optimal alignment that maximizes the similarity between source and target sentences. We modified the sentence alignment toolkit [11] avail-

[1] http://www.dmoz.org

Corpus	**Recursion depth**	Bilingual webpages	sentences	words	
				en	es
Web	depth 0	20186	176816	1597588	1729584
	depth 1	125027	1120762	8589037	9214945
	depth 2	1072957	3506675	26851681	28459003
Europarl	N/A	N/A	1486101	37184450	38996284

Table 1: Statistics of the data used in experiments. Depth 0 refers to links generated by the bilingual Web crawling strategy and depth 1, depth 2 refer to the links obtained by recursive intra-site crawling

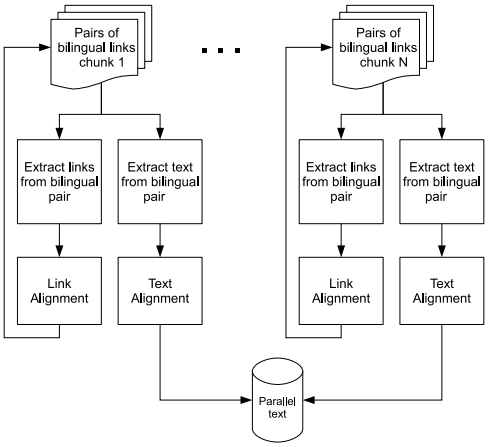

Figure 1: Illustration of the recursive intra-site mining approach. The parallelization offered by our approach is shown through the chunks.

able through Linguistic Data Consortium (LDC) to perform the alignment. The modified tool accepts a list containing pairs of root Web sites and performs the alignment of segmented text as well as links using a bilingual dictionary. The dictionary in our experiments was obtained by consolidating the forward and inverse word alignment model trained on Europarl corpus [12] using GIZA++ [13]. The 5-best translations for each source word (the dictionary obtained using word alignment is probabilistic) were included in the bilingual dictionary for sentence alignment. We also experimented with a manually created dictionary (see Section 4). During the text acquisition phase from a pair of bilingual links, we use the bilingual dictionary to compute a page level cosine metric. Only pairs of pages with cosine metric greater than 0.7 were retained for the sentence alignment phase. We accelerated the alignment process by keeping the bilingual dictionary in memory throughout the alignment process and recursively mined the root bilingual sites. The entire procedure was parallelized.

In each step of the recursion, we retain only those links that have the same domain as the root from which it originated. Such a procedure ensures that random Web sites that may be present as external links are not crawled. To enable link alignment using dynamic programming, we tokenized the link addresses and removed all the punctuation. We used a simple filter to separate compound words joined with "and" ("y") and "or" ("o"). For example, the address {http://www.eivissaweb.com/directory/artand culture/} is transduced into {http www eivissaweb com directory art and culture}. We also added a few rules to the bilingual dictionary such as english $\rightarrow$ espanol, en $\rightarrow$ es, eng $\rightarrow$ esp, etc. and enabled identity matches for link alignment.

4. Machine Translation Experiments

In this section, we exploit the parallel text obtained through our crawling strategy as augmented data in machine translation. We use a phrase-based statistical machine translation system [14] in all the experiments.

4.1. Data

In this work, the domain we are interested in is tourism and hospitality services for English and Spanish. The initial set of 20186 link pairs generated by the bilingual Web crawler was recursively mined using the method described in Section 3. We used a depth of 2 for the recursion. The statistics of the data obtained at each level of recursion is shown in Table 1. In many bilingual page pairs, the headers, footers as well as some menu items may appear in the same language. Further, the alignment process may also generate some errors. We use a simple length and word-based filtering approach similar to [4] to eliminate anomalous sentence pairs obtained from the crawling. The filter picks only those sentence pairs with sentence length ratio less than two and ensure that at least half the words in the source sentence have a translation in the target sentence, according to the bilingual dictionary. We also retain only those pairs that contain source sentences that have more than 70% of words in the source language. The filtered set contains 2,039,272 bilingual sentence pairs.

In order to obtain a representative reference development and test set, we manually picked two pairs of bilingual Web sites: {http://www.usatourist.com/english/index.html, http://www.usatourist.com/espanol/index.html} and {http://www.spain.info, http://www.spain.info/es/}. Both the bilingual Web site roots have a systematic hyperlink structure that can be easily aligned using our dynamic programming algorithm. We used our recursive mining technique ($recursion\ depth\ =\ 5$) to harvest bilingual text from these Web sites. This resulted in a set of about 10,000 unique sentence pairs that were manually verified by a bilingual (English-Spanish) speaker. The final set comprised a set of 7100 sentence pairs, 1000 of which were randomly chosen as a development set.

4.2. Results

We performed machine translation experiments in both directions, English-Spanish and Spanish-English. The baseline model was trained on Europarl [12] corpus. The model can be considered to be out-of-domain with respect to our test domain. The Web data translation model was trained on the 2,039,272 bilingual sentences extracted using our scheme. We also used a combination of the two models that we call as combined model. The combined model uses both the phrase tables during decod-

Model	Training data	English-Spanish			Spanish-English		
		BLEU	METEOR	TER	BLEU	METEOR	TER
Baseline	Europarl	20.96	21.21	60.65	22.36	48.37	64.65
	Web	25.06	22.80	56.86	28.13	52.75	58.00
	Combined	26.26	23.68	55.65	28.68	53.29	57.19
MERT	Europarl	22.76	20.85	62.13	25.82	50.22	58.04
	Web	26.19	22.52	58.23	30.69	52.87	52.66
	Combined	27.65	22.96	57.52	32.73	54.18	50.58

Table 2: Automatic evaluation metric scores for translation models from out-of-domain data, in-domain Web data and combined models.

ing. The reordering table was also concatenated from the two models. Table 2 presents the translation performance in terms of various metrics such as BLEU [15], METEOR [16] and Translation Edit Rate (TER) [17]. The results are presented for baseline models as well as models optimized by using Minimum Error Rate Training (MERT) [18] on the development set. In both cases, the language model was a 5 gram language model optimized on the development set based on perplexity. For the baseline model, the translation weights are uniform and for the optimized model, the weights of the log-linear model are learned using MERT.

While the out-of-domain model trained using Europarl data achieves a BLEU score of 22.76 on the test set (tourism and hospitality domain) for English-Spanish, the model constructed from our bilingual crawl achieves a 15.0% relative improvement. The interpolated model achieves an additional 6.5% improvement. Similar improvements hold for Spanish-English translation. For all three objective metrics, we achieve significant improvements in translation performance. The Web data also aids in reducing the out-of-vocabulary (OOV) rate of the translation models. The OOV rate of source side for English-to-Spanish translation drops from 22.8% to 8.9% when the Web data is added to the Europarl data. The target side OOV also drops from 21.8% to 8.4%. Clearly, the data from the Web crawling is beneficial in learning translation of words not seen in the seed corpus.

5. Conclusions and Future Work

We presented a new bilingual Web crawling strategy that focuses on locating bilingual Web sites. Our bilingual crawling strategy produces high quality entry points into bilingual Web sites that are subsequently mined through a novel recursive mining technique. The recursive mining technique uses dynamic programming to align both the text and links obtained from a pair of webpages. The recursive mining technique resulted in a set of 1.3 million Web page pairs from an initial set of 20186 link pairs obtained from the bilingual crawler, yielding approximately 4.7 million sentence pairs. Filtering the parallel text by using sentence similarity measures (through a word-based lexicon) resulted in about 2 million high quality parallel sentences. Our domain of interest in this work was tourism and hospitality services. Augmenting machine translation model with the in-domain Web text resulted in a relative improvement of 21% in BLEU score over an out-of-domain seed model trained on European parliamentary proceedings. We are currently working on modifying the visitation policy of the crawler to obtain data for other domains (news, health, etc.). We are also experimenting with different language pairs beyond English and Spanish.

6. References

[1] X. Zhu and R. Rosenfield, "Improving trigram language modeling with the world wide web," in *Proceedings of ICASSP*, 2001.

[2] Y. Tsvetkov and S. Wintner, "Automatic acquisition of parallel corpora from websites with dynamic content," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2010.

[3] P. Resnik and N. A. Smith, "The web as a parallel corpus," *Computational Linguistics*, vol. 29, pp. 349–380, September 2003.

[4] D. S. Munteanu and D. Marcu, "Improving machine translation performance by exploiting non-parallel corpora," *Comput. Linguist.*, vol. 31, pp. 477–504, December 2005.

[5] J. Uszkoreit, J. M. Ponte, A. C. Popat, and M. Dubiner, "Large scale parallel document mining for machine translation," in *Proceedings of COLING*, 2010, pp. 1101–1109.

[6] Y. Zhang, K. Wu, J. Gao, and P. Vines, "Automatic Acquisition of Chinese–English Parallel Corpus from the Web," *Advances in Information Retrieval*, pp. 420–431, 2006.

[7] L. Barbosa, S. Srinivas Bangalore, and V. K. Rangarajan Sridhar, "Crawling back and forth: Using back and out links to locate bilingual sites," *Submitted*, 2011.

[8] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the Web for emerging cyber-communities," *Computer networks*, vol. 31, no. 11-16, pp. 1481–1493, 1999.

[9] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Proceedings of Workshop on Artificial Intelligence for Web Search*, 2000.

[10] N. K. Gupta and S. Bangalore, "Segmenting spoken language utterances into clauses for semantic classification," in *Proceedings of ASRU*, 2003.

[11] X. Ma, "Champollion: A robust parallel text sentence aligner," in *LREC 2006: Fifth International Conference on Language Resources and Evaluation*, Genova, Italy, 2006.

[12] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *MT Summit*, 2005.

[13] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[14] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: open source toolkit for statistical machine translation," in *Proceedings of ACL*, 2007, pp. 177–180.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002.

[16] A. Lavie and M. Denkowski, "The METEOR metric for automatic evaluation of machine translation," *Machine Translation*, 2010.

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of AMTA*, 2006.

[18] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of ACL*, 2003.