Finding More Bilingual Webpages with High Credibility via Link Analysis

Chengzhi Zhang*
Nanjing University of Science and Technology
Nanjing, China

Xuchen Yao†
Johns Hopkins University
Baltimore, MD, USA

Chunyu Kit
City University of Hong Kong, Hong Kong SAR, China

Abstract

This paper presents an efficient approach to finding more bilingual webpage pairs with high credibility via link analysis, using little prior knowledge or heuristics. It extends from a previous algorithm that takes the number of bilingual URL pairs that a key (i.e., a URL pairing pattern) can match as the objective function to search for the best set of keys yielding the greatest number of webpage pairs within targeted bilingual websites. Enhanced algorithms are proposed to match more bilingual webpages following the credibility based on statistical analysis of the link relationship of the seed websites available. With about 12,800 seed websites as test set, the enhanced algorithms improve precision over baseline by more than 5%, from 94.06% to 99.40%, and hence find above 20% more true bilingual URL pairs, illustrating that significantly more bilingual webpages with high credibility can be mined with the help of the link analysis.

1 Introduction

Parallel corpora of bilingual text (bitext) are indispensable language resources for many data-driven tasks of natural language processing, such as statistical machine translation (Brown et al., 1990), cross-language information retrieval (Davis and Dunning, 1995; Oard, 1997), and bilingual lexical acquisition (Gale and Church, 1991; Melamed, 1997; Jiang et al., 2009), to name but a few. A general way to develop such corpora from web texts starts from exploring the structure of known bilingual websites, which are usually organized by their web masters in a way to facilitate both navigation and maintenance (Nie, 2010). The most common strategy is to create a parallel structure in terms of URL hierarchies, exploiting some known naming conventions for webpages of corresponding languages (Huang and Tilley, 2001; Nie, 2010). Following available structures and naming conventions, researchers have been exploring various means to mine parallel corpora from the web and a good number of such systems have demonstrated the feasibility and practicality in automatic acquisition of parallel corpora from bilingual and/or multilingual web sites, e.g., STRAND (Resnik, 1998; Resnik, 1999; Resnik and Smith, 2003), BITS (Ma and Liberman, 1999), PTMiner (Chen and Nie, 2000), PTI (Chen et al., 2004), WPDE (Zhang et al., 2006), the DOM tree alignment model (Shi et al., 2006), PagePairGetter (YE et al., 2008) and Bitextor (Esplà-Gomis and Forcada, 2010).

Most of these systems are run in three steps: first, bilingual websites are identified and crawled; second, pairs of parallel webpages are extracted; and finally, the extracted pairs are validated (Kit and Ng, 2007). Among them, prior knowledge about parallel webpages, mostly in the form of ad hoc heuristics for identifying webpage languages or pre-defined patterns for matching or computing similarity between webpages, is commonly used for webpage pair extraction (Chen and Nie, 2000; Resnik and Smith, 2003; Zhang et al., 2006; Shi et al., 2006; Yulia and Shuly, 2010; Tomás et al., 2008). Specifically, these systems exploit search engines and heuristics across webpage anchors to locate candidate bilingual websites and then identify webpage pairs based on pre-defined URL matching patterns. However, ad hoc heuristics cannot exhaust all possible patterns. Many webpages do not even have any language label in their anchors, not to mention many untrustworthy labels. Also, using a limited set of pre-

*Performed while a research associate at City University of Hong Kong.

†Performed while a visiting student at City University of Hong Kong.

Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 138–143,
Sofia, Bulgaria, August 8, 2013.

defined URL patterns inevitably means to give up all reachable bilingual webpages that fall outside their coverage.

Addressing such weaknesses of the previous approaches, we instead present an efficient bilingual web mining system based on analyzing link relationship of websites without resorting to prior ad hoc knowledge. This approach extends, on top of re-engineering, the previous work of Kit and Ng (2007). It aims at (1) further advancing the idea of finding bilingual webpages via automatic discovery of non-ad-hoc bilingual URL pairing patterns, (2) applying the found pairing patterns to dig out more bilingual webpage pairs, especially those involving a deep webpage unaccessible by web crawling, (3) discovering more bilingual websites (and then more bilingual webpages) with high credibility via statistical analysis of bilingual URL patterns and link relationship of available seed websites. The results from our experiments on $12,800$ seed websites show that the proposed algorithms can find considerably more bilingual webpage pairs on top of the baseline, achieving a significant improvement of pairing precision by more than 5%.

2 Algorithm

This section first introduces the idea of unsupervised detection of bilingual URL pairing patterns (§2.1) and then continues to formulate the use of the detected patterns to explore more websites, including deep webpages (§2.2), and those not included in our initial website list (§2.3).

2.1 Bilingual URL Pattern Detection

Our current research is conducted on top of the re-implementation of the intelligent web agent to automatically identify bilingual URL pairing patterns as described in Kit and Ng (2007). The underlying assumption for this approach is that rather than random matching, parallel webpages have static pairing patterns assigned by web masters for engineering purpose and these patterns are put in use to match as many pairs of URLs as possible within the same domain. Given a URL u from the set U of URLs of the same domain, the web agent goes through the set $U-\{u\}$ of all other URLs and finds among them all those that differ from u by a single token[1] – a token is naturally separated by a special set of characters including slash /, dot ., hyphen -, and underscore _ in a URL. Then, the single-token difference of a candidate URL pairs is taken as a candidate of URL paring pattern, and all candidate patterns are put in competition against each other in a way to allow a stronger one (that matches more candidate URL pairs) to win over a weaker one (that matches fewer). For instance, the candidate pattern <en, zh> can be detected from the following candidate URL pair:

www.legco.gov.hk/yr99-00/en/fc/esc/e0.htm
www.legco.gov.hk/yr99-00/zh/fc/esc/e0.htm

The re-implementation has achieved a number of improvements on the original algorithm through re-engineering, including the following major ones.

1. It is enhanced from token-based to character-based URL matching. Thus, more general patterns, such as <e, c>, can be aggregated from a number of weaker ones like <1e, 1c>, <2e, 2c>, ..., etc., many of which may otherwise fail to survive the competition.
2. The original algorithm is speeded up from $O(|U|^2)$ to $O(|U|)$ time, by building inverted indices for URLs and establishing constant lookup time for shortest matching URL strings.[2]
3. The language detection component has been expanded from bilingual to multi-lingual and hence had the capacity to practically handle multilingual websites such as those from EU and UN.

When detected URL patterns are used to match URLs in a web domain for identifying bilingual webpages, noisy patterns (most of which are presumably weak keys) would better be filtered out. A straightforward strategy to do this is by thresholding the credibility of a pattern, which can be defined as

$$C(p, w) = \frac{N(p, w)}{|w|}.$$

where $N(p, w)$ is the number of webpages matched into pairs by pattern p within website w, and $|w|$ the size of w in number of webpages. Note that this is the *local* credibility of a key with respect to a certain website w. Empirically, Kit and

[1] If language identification has been done on webpages, it only needs to go through all URLs of the other language.

[2] Achieved by utilizing SecondString http://secondstring.sf.net/

Ng (2007) set a threshold of 0.1 to rule out weak noisy keys.

Some patterns happen to generalize across domains. The *global* credibility of such a pattern p is thus computed by summing over all websites involved, in a way that each webpage matched by p is counted in respect to the local credibility of p in the respective website:

$$C(p) = \sum_{w} C(p, w)\, N(p, w).$$

Interestingly, it is observed that many weak keys ruled out by the threshold 0.1 are in fact good patterns with a nice global credibility value. In practice, it is important to "rescue" a local weak key with strong global credibility. A common practice is to do it straightforwardly with a global credibility threshold, e.g., $C(p) > 500$ as for the current work.

Finally, the bilingual credibility of a website is defined as

$$C(w) = \max_{p} C(p, w).$$

It will be used to measure the bilingual degree of a website in a later phase of our work, for which an assumption is that bilingual websites tend to link with other bilingual websites.

2.2 Deep Webpage Recovery

Some websites contain webpages that cannot be crawled by search engines. These webpages do not "exist" until they are created dynamically as the result of a specific search, mostly triggered by JavaScript or Flash actions. This kind of webpages as a whole is called *deep web*. Specifically, we are interested in the case where webpages in one language are visible but their counterparts in the other language are hidden. A very chance that we may have to unearth these deep hidden webpages is that their URLs follow some common naming conventions for convenience of pairing with their visible counterparts.

Thus for each of those URLs still missing a paired URL after the URL matching using our bilingual URL pattern collection, a candidate URL will be automatically generated with each applicable pattern in the collection for a trial to access its possibly hidden counterpart. If found, then mark them as a candidate pair. For example, the pattern `<english,tc_chi>` is found applicable to the first URL in Table 1 and accordingly generates the second as a candidate link to its English counterpart, which turns out to be a valid page.

2.3 Incremental Bilingual Website Exploration

Starting with a seed bilingual website list of size N, bilingual URL pairing patterns are first mined, and then used to reach out for other bilingual websites. The assumption for this phase of work is that bilingual websites are more likely to be referenced by other bilingual websites. Accordingly, a weighted version of PageRank is formulated for prediction.

Firstly, outgoing links and PageRank are used as baselines. *Linkout*(w) is the total number of outgoing links from website w, and the PageRank of w is defined as (Brin and Page, 1998):

$$\textit{PageRank}(w) = \frac{r}{N} + (1-r) \sum_{w \in M(w)} \frac{\textit{PageRank}(w)}{\textit{Linkout}(w)},$$

where $M(w)$ is the set of websites that link to w in the seed set of N bilingual websites, and $r \in [0, 1]$ a damping factor empirically set to 0.15. Initially, the *PageRank* value of w is 1. In order to reduce time and space cost, both *Linkout*(w) and *PageRank*(w) are computed only in terms of the relationship of bilingual websites in the seed set.

The *WeightedPageRank*(w) is defined as the *PageRank*(w) weighted by w's credibility $C(w)$. To reach out for a related website s outside the initial seed set of websites, our approach first finds the set $R(s)$ of seed websites that have outgoing links to s, and then computes the sum of these three values over each outgoing link, namely, $\sum_w$*Linkout*(w), $\sum_w$*PageRank*(w), and $\sum_w$*WeightedPageRank*(w) for each $w \in R(s)$, for the purpose of measuring how "likely" s is bilingual. An illustration of link relationship of this kind is presented in Figure 1.

In practice, the exploration of related websites can be combined with bilingual URL pattern detection to literately harvest both bilingual websites and URL patterns, e.g., through the following procedure:

1. Starting from a seed set of websites as the current set, detect bilingual URL patterns and then use them to identify their bilingual webpages.
2. Select the top K linked websites from the seed set according to either $\sum$*Linkout*, $\sum$*PageRank*, or $\sum$*WeightedPageRank*.

(1) http://www.fehd.gov.hk/tc_chi/LLB_web/cagenda_20070904.htm
(2) http://www.fehd.gov.hk/english/LLB_web/cagenda_20070904.htm

Table 1: Illustration of URL generation for a deep webpage

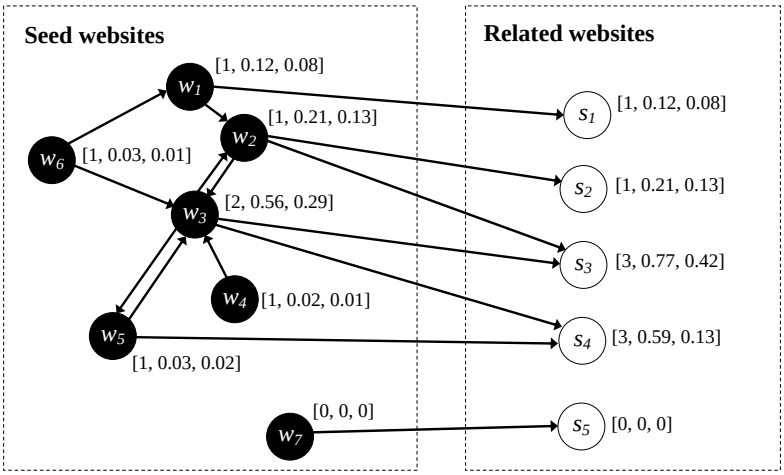

Figure 1: Illustration of link relationship of seed websites and related websites, with associated $\sum$*Linkout*, $\sum$*PageRank* and $\sum$*WeightedPageRank* in square brackets and with arrows to indicate outgoing links from a seed website to others.

3. Add the top K selected websites to the current set, and repeat the above steps for desired iterations.

3 Evaluation

The implementation of our method results in PupSniffer,[3] a Java-based tool that has been released for free. A series of experiments were conducted with it to investigate the performance of the proposed method on about $12,800$ seed websites. A web interface was also implemented for evaluating the candidate bilingual webpage pairs identified by our system.

3.1 Seed Websites

The initial seed websites were collected from two resources, namely

- Hong Kong Website Directory[4] and
- Hong Kong World Wide Web Database.[5]

After the removal of invalid ones, $12,800$ websites were finally acquired as our seed set.[6]

3.2 URL Pattern Detection and Deep Webpage Recovery

The enhanced algorithm described in Section 2.1 above was ran to extract credible URL patterns. In general, the extracted patterns are valid as long as the threshold is not too low – it is set to $C(p, w) > 0.1$ in our experiments. A number of strongest patterns found are presented in Table 2 for demonstration. Most of them, especially <en, tc> and <eng, chi>, are very intuitive patterns. A full list of URL pairing patterns detected in our experiments is also available.[7] Particularly interesting is that all these patterns were identified in an unsupervised fashion without any manual heuristics.

Using these patterns, the original algorithm retrieved about $290K$ candidate bilingual webpage pairs. By the simple trick of rescuing weak local patterns with the global credibility threshold $C(p) > 500$, $10K$ more webpage pairs were further found. Additionally, other $16K$ webpage pairs were dug out from deep webpages by automatically generating paired webpages with the aid of identified URL patterns.

[3] http://code.google.com/p/pupsniffer

[4] http://www.852.com

[5] http://www.cuhk.edu.hk/hkwww.htm

[6] http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/Data/All_Seed_Websites_List.txt

[7] http://mega.ctl.cityu.edu.hk/~czhang22/pupsniffer-eval/Data/Pattern_Credibility_LargeThan100.txt

Pattern	$C(p)$
<en, tc>	13997.36
<eng, tc>	12869.56
<english, tc_chi>	11436.12
<english, chinese>	11032.46
<eng, chi>	7824.86

Table 2: Top 5 patterns with their global credibility values.

Method	Pairs	Precision
Kit and Ng (2007)	290,247	94.06%
Weak key rescue	10,015	89.27%
Deep page recovery	15,825	95.02%
Incremental exploration	37,491	99.40%
Total	348,058	94.72%
True pair increment	55,674	20.76%

Table 3: Number of bilingual webpage pairs found and their precision from sampled evaluation.

3.3 Website Exploration

To go beyond the original 12,800 websites, the incremental algorithm described in Section 2.3 was run for one iteration to find outside bilingual websites directly linked from the seeds. The top 500 of them, ranked by $\sum$*Linkout*, $\sum$*PageRank* and $\sum$*WeightedPageRank*, respectively, were manually checked by five students, giving the curves of the total number of true bilingual websites and overall precision per top N websites as plotted in Figure 2. These results show that almost 50% of the top 500 related outside websites ranked by $\sum$*WeightedPageRank* are true bilingual websites. A higher precision indicates more bilingual webpage pairs correctly matched by the URL patterns in use.

After one iteration of the incremental algorithm, $37K$ more candidate bilingual webpage pairs were found in the related outside websites, besides the $290K$ by the original algorithm. Table 3 presents the number of webpage pairs identified by each algorithm with a respective precision drawn from random sampling. These results suggest that our proposed enhancement is able to harvest above 20% more bilingual webpage pairs without degrading the overall precision. Error analysis shows that around 80% of errors were due to mistakes in language identification for webpages. For instance, some Japanese webpages were mistakenly recognized as Chinese ones.

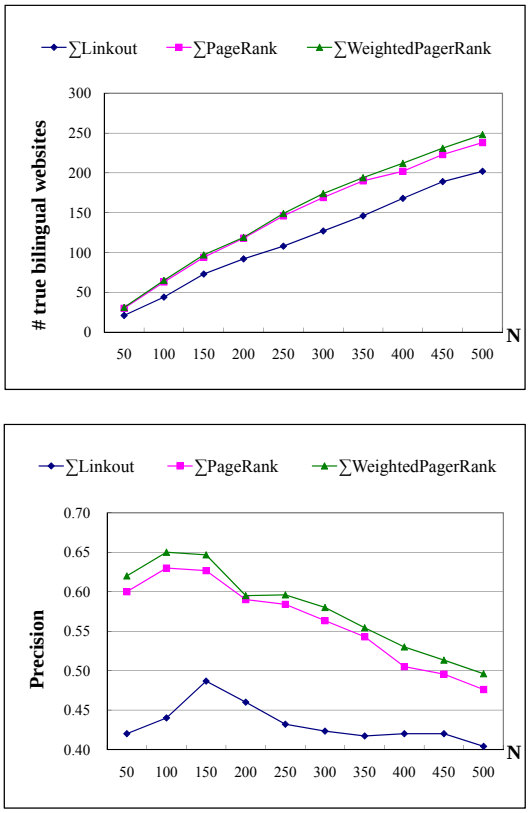

Figure 2: Number and precision of true bilingual websites found per top N outside websites ranked by various criteria.

4 Conclusion

In this paper we have presented an efficient approach to mining bilingual webpages via computing highly credible bilingual URL pairing patterns. With the aid of these patterns learned in an unsupervised way, our research moves on to exploring the possibility of rescuing weak local keys by virtue of global credibility, uncovering deep bilingual webpages by generating candidate URLs using available keys, and also developing an incremental algorithm for mining more bilingual websites that are linked from the known bilingual websites in our seed set. Experimental results show that these several enhanced algorithms improve the precision over the baseline from 94.06% to 99.40% and, more importantly, help discover above 20% more webpage pairs while maintaining a high overall precision.

Acknowledgements

The research described in this paper was supported in part by the Research Grants Council (RGC) of Hong Kong SAR, China, through the GRF

grant 9041597 (CityU 144410), National Natural Science Foundation of China through the grant No. 70903032, and Project of the Education Ministry of China's Humanities and Social Sciences through the grant No. 13YJA870020.

References

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Jiang Chen and Jian-Yun Nie. 2000. Parallel web text mining for cross-language ir. In *Proc. of RIAO*, pages 62–77.

Jisong Chen, Rowena Chau, and Chung-Hsing Yeh. 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*, pages 157–161.

Mark W Davis and Ted E Dunning. 1995. A trec evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference*, pages 483–498.

Miquel Esplà-Gomis and Mikel L Forcada. 2010. Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(1):77–86.

William A Gale and Kenneth W Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, pages 152–157.

Shihong Huang and Scott Tilley. 2001. Issues of content and structure for a multilingual web site. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 103–110.

Long Jiang, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 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*, pages 870–878.

Chunyu Kit and Jessica Yee Ha Ng. 2007. An intelligent web agent to mine bilingual parallel pages via automatic discovery of URL pairing patterns. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops: Workshop on Agents and Data Mining Interaction (ADMI-07)*, pages 526–529.

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

I. Dan Melamed. 1997. A word-to-word model of translational equivalence. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, pages 490–497.

Jian-Yun Nie. 2010. *Cross-Language Information Retrieval*. Morgan and Claypool Publishers.

Douglas W Oard. 1997. Cross-language text retrieval research in the USA. In *Proceedings of the Third DELOS Workshop: Cross-Language Information Retrieval*, pages 7–16.

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

Philip Resnik. 1998. Parallel strands: A preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber, and E. Hovy, editors, *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98)*, pages 72–82.

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.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 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*, pages 489–496.

Jesús Tomás, Jordi Bataller, Francisco Casacuberta, and Jaime Lloret. 2008. Mining Wikipedia as a parallel and comparable corpus. In *Language Forum*, volume 34.

Sha-ni YE, Ya-juan LV, Yun Huang, and Qun Liu. 2008. Automatic parallel sentences extraction from web. *Journal of Chinese Information Processing*, 22:67–73.

T Yulia and W Shuly. 2010. Automatic acquisition of parallel corpora from website with dynamic content. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pages 3389–3392.

Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic acquisition of chinese–english parallel corpus from the web. In *Advances in Information Retrieval*, pages 420–431. Springer.