

Assembling a parallel corpus from RSS news feeds

John Fry

Artificial Intelligence Center
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493 USA
fry@ai.sri.com

Linguistics Department
San José State University
One Washington Square
San José, CA 95192-0093 USA
jfry@email.sjsu.edu

Abstract

We describe our use of RSS news feeds to quickly assemble a parallel English-Japanese corpus. Our method is simpler than other web mining approaches, and it produces a parallel corpus whose quality, quantity, and rate of growth are stable and predictable.

1 Motivation

A parallel corpus is an indispensable resource for work in machine translation and other multilingual NLP tasks. For some language pairs (e.g., English-French) data are plentiful. For most language pairs, however, parallel corpora are either nonexistent or not publicly available.

The need for parallel corpora led to the development of software for automatically discovering parallel text on the World Wide Web. Examples of such web mining systems include BITS (Ma and Liberman, 1999), PTMiner (Chen and Nie, 2000), and STRAND (Resnik and Smith, 2003).

These web mining systems, while extremely useful, do have a few drawbacks:

- They rely on a random walk through the WWW (through search engines or web spiders), which means that the quantity and, more importantly, the quality of the final results are unpredictable.
- Their ‘generate-and-test’ approaches are slow and inefficient. For example, STRAND and PTMiner work by applying sets of hand-crafted substitution rules (e.g., `english` \rightarrow `big5`) to all candidate URLs and then checking those new URLs for content, while the BITS system considers the full cross product of web pages on each site as possible translation pairs.
- They sometimes misidentify web page pairs as translations when in fact they are not.

- Good translation pairs are often missed. STRAND, for example, reports recall scores of 60% for some language pairs (Resnik and Smith, 2003).
- Although their source code has not been published, some web mining systems appear to be quite complex to implement, requiring hand-crafted URL manipulation rules and expertise in HTML/XML, similarity scoring, web spiders, and machine learning.

This paper describes a much simpler approach to web mining that avoids these disadvantages. We used our method to quickly assemble an English-Japanese parallel corpus whose quality, quantity, and rate of growth are stable and predictable, obviating the need for quality control by bilingual human experts.

2 Approach

Our approach exploits two recent trends in the delivery of news over the WWW.

The first trend is the growing practice of multinational news organizations to publish the same content in multiple languages across a network of online news sites. Among the most prolific are online news sites in the domain of information technology (IT). For example, CNET Networks (<http://www.cnetnetworks.com>), a large IT media company, publishes stories in Chinese, English, French, German, Italian, Japanese, Korean, and Russian on its worldwide network of IT news sites. Another conglomerate, JupiterMedia (<http://www.jupiterweb.com>), publishes in English, German, Japanese, Korean, and Turkish.

The second trend is the use of RSS, an XML-based syndication format. RSS is increasingly used by both mainstream news web sites (e.g., wired.com, news.yahoo.com, and the IT news sites mentioned above) as well as sites that provide news-like content (e.g., slashdot.org and

Listing 1: Procmail code for creating our corpus

```
1 # .procmailrc file: extracts
2 # parallel URLs from RSS feeds
3 :0 HB
4 * ^User-Agent: rss2email
5 |url='grep -o http:.*'\
6 ;wget -O - $url\
7 |egrep \
8 '(English)|CNET Networks|
9 target=original|<I>.*N</I></A>'\
10 |grep -o http:.*\
11 |sed -e 's/[?"].*//'\
12 |xargs -r echo -e "$url\t"\
13 >>parallel_url_list.txt
```

weblogs). RSS-aware client programs, called news aggregators, help readers keep up with such sites by displaying the latest headlines as soon as they are published. In other words, readers subscribe to the sites' RSS feeds, rather than checking the sites manually for new content.

In many cases, a story published in a target language (say Japanese) will include a link to the original story in the source language (usually English). When the target articles are published over RSS, as they increasingly are, then virtually all the ingredients of a parallel corpus are in place, with no random crawling required.

3 Assembling the parallel corpus

Using RSS feeds in the domain of technology news, we were able to automatically assemble an English-Japanese parallel corpus quickly with little programming effort.

The first step in assembling our corpus was to find web sites that publish Japanese-language news stories along with links to the original source articles in English. Table 1 lists the four RSS feeds we subscribed to. Instead of using a news aggregator, we subscribed to the sites in Table 1 using the open-source `rss2email` program (written by Aaron Swartz), which delivers news feed updates over email.

We then relied on standard UNIX tools like `procmail`, `grep`, `sed`, and `wget` to process the incoming RSS feeds as they arrived by email.

Listing 1 shows our `.procmailrc` configuration file that instructs `procmail` how to process incoming RSS feeds. First, the URL of the new Japanese story is extracted from the email (line 5), and the article is downloaded (line 6). Next, the link (if any) to the English source article is extracted from the Japanese article (lines 7-11).

Finally, both the Japanese and English URLs are saved to a file (lines 12-13).

The regular expression in lines 8-9 of Listing 1 matches text that accompanies a link to the English source article. This is the only part of Listing 1 that is specific to the sites we used (Table 1) and that would need to be modified in order to adapt our method to different languages or web sites.

It should be noted that we do not record the *content* of the parallel news articles. Because material on the Web is subject to copyright restrictions, we cannot publish the content directly. Rather, we record the *URL* of each pair of Japanese and English articles, separated by a tab character. This same format, tab-separated URLs, is also used by the STRAND project for distributing their web-mined parallel corpora (Resnik and Smith, 2003). The STRAND web page (<http://umiacs.umd.edu/~resnik/strand>) offers a short Perl program for extracting the actual content from the URL pairs; this program works for our English-Japanese data as well.

4 Results

4.1 A five-week RSS corpus

We processed the RSS feeds from the Japanese sources listed in Table 1 over a period of five weeks. At the end of the fifth week, we had collected 333 parallel article pairs in Japanese and English. As Figure 1 shows, the bulk of the 333 article pairs were collected from HotWired (133) and CNET (125), followed by pairs from internet.com (65) and IT Media (10).

We then manually inspected all 333 translation pairs to check for problems. One of the URLs we collected, ostensibly a link to an English-language CNET article, turned out to be a stale link. Another link, from a Japanese internet.com article, did not in fact point to an English translation. Finally, two of the HotWired translation pairs were found to be repeats (HotWired occasionally republishes popular articles from the past). We discarded the two bad pairs and the two repeats, leaving us with a final corpus of 329 unique translation pairs.

4.2 Supplementing the corpus by crawling the archives

A corpus of 329 parallel news articles is of course insufficient for most tasks. We therefore recursively crawled the past archives of all four

URL of main news site	RSS feed
http://hotwired.goo.ne.jp	http://www.hotwired.co.jp/news/index.rdf
http://japan.cnet.com	http://japan.cnet.com/rss
http://japan.internet.com	http://bulknews.net/rss/rdf.cgi?InternetCom
http://www.itmedia.co.jp	http://bulknews.net/rss/rdf.cgi?ITmedia

Table 1: RSS feeds used to construct our parallel English-Japanese corpus

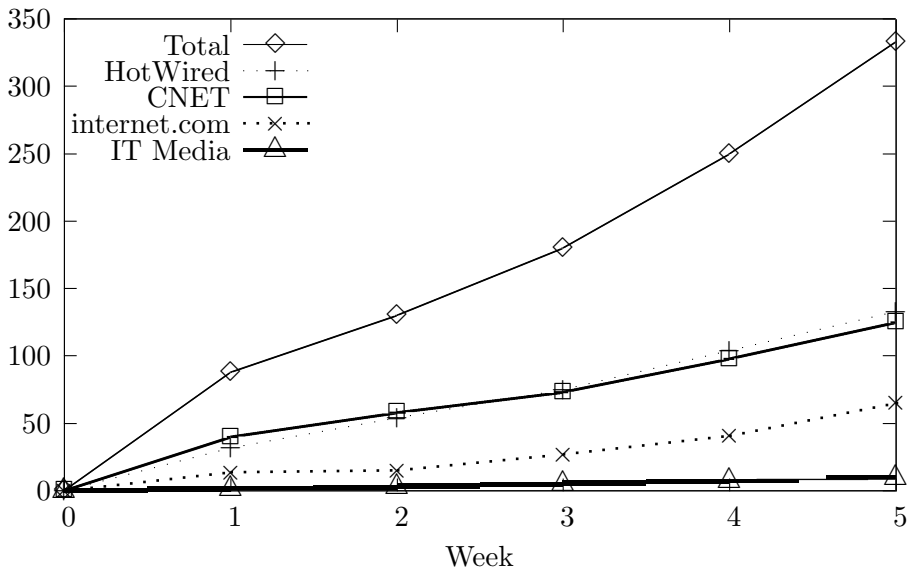


Figure 1: Sources of the 333 parallel English-Japanese article pairs collected over five weeks

HotWired	6,701
CNET	328
internet.com	8,227
ITMedia	2,021
Total	17,277

Table 2: Total article pairs after crawling

web sites in Table 1 using `wget -r` to find article pairs that were posted earlier than those in our five-week experiment with RSS feeds. This crawling netted more than 15,000 additional article pairs. The total count of collected article pairs as of the time of writing (duplicates removed) is shown in Table 2.

As Table 2 shows, the bulk of the article obtained through crawling came from the HotWired and internet.com sites, both of which maintain archives stretching back several years. ITMedia, on the other hand, observes a policy of removing content after one year, and so is not a substantial source for archived material.

4.3 Availability of our data

A regularly updated list of all English-Japanese article pairs we have collected so far can

be downloaded from http://johnfry.org/je_corpus. At the time of writing, the list holds 17,277 English-Japanese article pairs (see Table 2), and is growing at a rate of approximately 70 pairs per week.

Where possible, our collected URLs point to ‘printer-friendly’ (as opposed to ‘screen-friendly’) versions of the content. Printer-friendly versions of news articles are structurally simpler, with fewer banners and advertisements cluttering the story content. In addition, printer-friendly versions typically contain the entire news article, whereas screen-friendly versions are sometimes published over several successive pages, making them more difficult to process. All the HotWired and IT Media articles in our corpus have printer-friendly versions in both English and Japanese. In the case of CNET and internet.com, the English-language articles offer printer-friendly versions, but the Japanese articles do not.

5 Other parallel English-Japanese corpora on the web

Our corpus of 17,277 article pairs is the largest, but not the only, parallel English-Japanese cor-

pus that is freely available on the web. The following are other free sources of English-Japanese data:

- The NTT Machine Translation Research Group offers a set of 3,718 Japanese-English sentence pairs at <http://www.kecl.ntt.co.jp/icl/mtg/resources>
- The OPUS project at <http://logos.uio.no/opus> offers 33,143 aligned Japanese-English sentence pairs, taken from the documentation for the OpenOffice software suite.
- A set of aligned translations of 114 works of literature (taken from Project Gutenberg and similar sources) is available from the homepage of NICT researcher Masao Utiyama at <http://www2.nict.go.jp/jt/a132/members/mutiyama>

A substantial, but not free, source of Japanese-English data is the set of 150,000 aligned sentence pairs collected from newspaper articles and aligned by Utiyama and Isahara (1993). This collection can be licensed from NICT (see the above link to Utiyama's home page).

6 Conclusion

We have demonstrated how RSS news feeds can be used to quickly assemble a parallel corpus. In the case of our Japanese-English corpus, we supplemented the RSS feeds with web crawling of the news archives in order to assemble a corpus of substantial size (17,277 article pairs and growing).

One drawback of our method is that it is feasible only for language pairs with a substantial online news media representation. On the other hand, our approach has two major advantages over web mining systems. First, it is considerably simpler to implement, requiring essentially one extended line of pipelined Unix shell commands (Listing 1). Second, our approach produces a parallel corpus whose quantity, rate of growth, and (most importantly) quality are stable and predictable. The burden of quality control (including article quality, translation quality, and identification of translation pairs) is shifted onto the news organization that publishes the RSS feed, rather than resting on the web crawling system or bilingual human experts.

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, Seattle.
- Xiaoyi Ma and Mark Liberman. 1999. BITS: a method for bilingual text search over the web. In *Proceedings of MT Summit VII*, September.
- Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.
- Masao Utiyama and Hitoshi Isahara. 1993. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of ACL-93*, pages 72–79, Columbus, Ohio.