Harvesting Wikipedia Interwiki Links

Darren Cook
darren@dcook.org

Introduction

Wikipedia [1] has what are called interwiki links, most commonly used to link an entry in one language to the corresponding entry in another language. For instance the English Wikipedia article for "Computer Science" [2] has interwiki links to a Japanese article called "計算機科学", a German article called "Informatik" and a Chinese article called "计算机科学". This suggests they could be harvested to make a dictionary.

Extraction Process

Automatic extraction is facilitated by the fact that Wikipedia freely makes available XML dumps [3] of all its content, though they can sometimes be a few months out of date. For our purposes just the title and the interwiki link are used; the body of the article is ignored. The title of the target page (in the target language) is conveniently already part of the interwiki link in the XML file, so there is no need to do any further lookup or cross-referencing with another XML dump.

The XML files are very large (as of Nov 2007 the English file is the largest at 3Gb compressed), so they were processed by a PHP script [4] using the built-in XML stream parser [5]. A tree-based XML parser [6] would need to hold the whole document in memory which is unreasonable. A stream parser, on the other hand, need only hold one Wikipedia article in memory at a time. By having the script read from STDIN and using a commandline, like the following, the compressed files can be used as-is, without need to first uncompress them (which requires considerable disk space):

```
bzip2 -dc pages_current.xml.bz2 | php parse.php en en,ja,de,zh,ar
```

This example command takes 30 to 60 minuets to run on recent hardware, and produces en_ar.txt, en_de.txt, en_en.txt, en_ja.txt, en_zh.txt files. A similar command was run on the Japanese wikipedia XML dump file to produce ja_ar.txt, ja_de.txt, ja_en.txt, ja_ja.txt and ja_zh.txt files. The process was repeated for German, Chinese and Arabic (all languages supported by Wikipedia are supported by the extraction script). Table One shows the number of extracted entries.

	AR	DE	EN	JA	ZH
AR	*8,584*	21,776	28,097	16,326	15,374
DE	21,350	*319,546*	342,425	101,053	55,113
EN	26,822	327,433	*1,563,191*	163,389	81,509
JA	15,415	102,165	173,995	*146,453*	58,425
ZH	14,805	57,640	89,802	60,332	*95,119*

(How to read: AR row is for the data read from the Arabic wikipedia; the columns show how many entries were found for that target language; so 16,326 links to Japanese articles were found in the Arabic Wikipedia; 15,415 links to Arabic articles were found in the Japanese Wikipedia.)
(The italicized numbers on the diagonal show number of REDIRECTs within the same language.)

Merge And Normalization Process

The files produced are two column tab-separated UTF-8 text files, and they are for one-way lookup. They could be used in this format, but another PHP script [7] is used to import them into an SQL database. The major feature of this import process is that we compare the two one-way lookup files and only keep entries that mirror each other; we then create a two-way table. There is also some clean-up and normalization: any entries containing brackets are thrown away, alphabet characters are lowercased, hex entities are converted. For non-European languages any entries containing just ASCII characters are thrown away. Mostly these are valid dates or abbreviations, but sometimes they are disambiguation pages or simply mistakes. We therefore lose some useful entries but also lose bad entries, so overall the accuracy of the dictionary improves. Table 2 shows the sizes of the language pairs. Import speed (on a 2.4Ghz quad-core machine importing to mysql on localhost) was about 65,000 entries per minute, so all ten language pairs shown could be processed and imported in about 15 minutes.

Language Pair	Word Pairs
ar-de	14,957
ar-en	20,997
ar-ja	11,214
ar-zh	9,430
de-en	261,071
de-ja	77,424
de-zh	41,833
en-ja	127,047
en-zh	63,224
ja-zh	45,976

Evaluation

The dictionaries were used to automatically match against WordNet to generate some initial data for MLSN for each language [9]. For the case of Japanese-English, using just jmdict [8] we could automatically translate 5,779 of WordNet's entries, with high or medium confidence. When we used both jmdict and the interwiki dictionary files that figure increased to 12,280 (and in this case the high and medium confidence categories are approximately 95% accurate on automatically translating WordNet, with blame for mistakes being split about equally between each dictionary source).

As an example of one of the cases where the interwiki link is responsible for a mistake, is 06microbrewery0 (that is its MLSN code). When using both jmdict and interwiki dictionaries together this ended up in the low confidence file but it is still an instructive example.

In WordNet the only entry in the synset is "microbrewery" and the gloss is "a small brewery; consumption of the product is mainly elsewhere". The English Wikipedia entry for microbrewery [10] has a pointer to the Japanese article called "地ビール" (jibiiru). However careful study of that page shows that the correct translation for microbrewery is "地ビールの醸造所", and that "地ビール" refers to the drink itself: "craft beer" or "microbrew". Incidentally English Wikipedia has no distinct entry for either of those terms: they redirect to the microbrewery entry. Similarly Japanese Wikipedia has no entry for 地ビールの醸造所. WordNet too does not have a "craft beer" entry.

In other words the problem here is that interwiki links only have to point to the closest corresponding article; they are not obliged to map exactly. Therefore it is important not to trust them completely and to always use them in conjunction with a second dictionary.

Just using the interwiki dictionary gives 12,823 high/medium hits, but quality has suffered: the different dictionary source kept it honest. For instance the "microbrewery" mistake described above ends up in the high confidence file instead of in the low confidence file. JMDict correctly has "local beer, microbrew" for 地ビール, but no entry for microbrewery. So, when using just JMDict, WordNet's microbrewery entry was not matched at all.

Similar Experiments

en_en.txt contains a list of potential English synonyms. These are found by looking for the REDIRECT strings in the XML file. These are not currently used by MLSN because the data is not very clean: they are a mix of genuine synonyms, spelling corrections and otherwise related words. (See the microbrewery discussion above for exactly this case: "craft beer" redirects to "microbrewery", which is a "part of" relation not a synonym relation.)

The same process was attempted on the Wiktionary [11] data. This was hoped to be even more useful as Wiktionary explicitly lists translations into other languages. Unfortunately this data is unstructured and not used consistently. Attempts to mine those translations soon got overwhelmed with a plethora of special cases and the attempt was abandoned.

Conclusion

We have shown how the knowledge captured in Wikipedia's interwiki links can be extracted to automatically generate bilingual dictionaries for all major world languages. The dictionaries are large, accurate and, because they have good coverage of non-dictionary words (products, companies, movies, famous people, etc.) they form a good complement to existing freely available dictionary sources. In addition informal comparisons to earlier experiments suggest the usefulness of the interwiki dictionary is increasing significantly as time goes on.

References And Links

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

[2]: http://en.wikipedia.org/wiki/Computer_science

[3]: http://en.wikipedia.org/wiki/Wikipedia:Database_download
http://download.wikimedia.org/

[4]: Part of the open-source fclib library (utilities/wikipedia_interwiki_extractor.php)
http://www.dcook.org/software/fclib/

[5]: PHP's XML Stream library, a wrapper for expat:
http://jp2.php.net/manual/en/ref.xml.php

[6]: Tree parsers built-in to PHP:
http://jp2.php.net/manual/en/ref.dom.php
http://jp2.php.net/manual/en/ref.simplexml.php

[7]: utilities/simpledict_import.php in fclib [4]

[8]: http://www.csse.monash.edu.au/~jwb/j_jmdict.html

[9]: Automatic Translation For MLSN, Cook. 2008.
http://dcook.org/mlsn/about/

[10]: http://en.wikipedia.org/wiki/Microbrewery

[11]: http://www.wiktionary.org/