Extracting bilingual word pairs from Wikipedia

Francis M. Tyers[*], **Jacques A. Pienaar**[†]

[*]Grup Transducens,
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant
03071 Alacant, Spain
ftyers@dlsi.ua.es

[*]Prompsit Language Engineering S.L.,
Pol. Ind. Canastell. Ctra. Agost, 77
03690 Sant Vicent del Raspeig, Spain

[†]Centre for Text Technology, North-West University
Potchefstroom 2531, South Africa
jacques.pienaar@nwu.ac.za

Abstract

A bilingual dictionary or word list is an important resource for many purposes, among them, machine translation. For many language pairs these are either non-existent, or very often unavailable owing to licensing restrictions. We describe a simple, fast and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the `interwiki` link system) and assess the performance of this method with respect to four language pairs. Precision was found to be in the 69–92% region, but open to improvement.

1. Introduction

Bilingual dictionaries are an important resource for natural language processing, for example cross-language information retrieval, and especially machine translation. In machine translation they are central to rule-based systems and useful in statistical machine translation (Koehn and Knight, 2002).

While bilingual dictionaries exist for pairs of larger languages such as English – French, they are scarce resources for many smaller language pairs. This includes pairs where a smaller language is paired with a larger language, for example English – Afrikaans.

Wikipedia is an online, collaboratively edited encyclopaedia with articles available in 256 languages (Wikipedia, 2008b). Neither the addition nor maintenance of Wikipedia entries requires any specific expertise over being able to use a standard web browser and entering text using a simple markup language. These encyclopaedias are freely-editable, and freely-distributable, which makes them the ideal platform for developing encyclopaedias in all the world's languages. The content and structure of these encyclopaedias make them amenable to linguistic research, whilst the breadth of language coverage makes them appropriate and useful for creating linguistic resources for languages which lack them.

For each language a separate encyclopaedia exists with its own norms and the articles between the different encyclopaedias are not simply translations of one another.

Each Wikipedia article can provide links to other articles on the same subject in different languages, so for example on the English article for *socialism* there is a link to the same article, *socijalizam*, on the Serbo-Croatian Wikipedia. These links are between article titles, which may be a word or a phrase.

The links between the same article in different languages are called `interwiki` links and are periodically maintained by bots.[1] This maintenance can occur in either *supervised* or *unsupervised* mode and is intended to keep the consistency of the links between the various Wikipedias. The general functionality of both modes of execution are the same. For each article the bot first checks the existing `interwiki` links of the *source* article. If any are found it then retrieves the articles they point to. The bot then adds links from the *target* articles which were not included in the *source* article, to the *source* article. If more than one link is retrieved in supervised mode, for any given language pair, then the operator of the bot is asked to pick the correct one, whilst in automatic mode, ambiguous links are skipped. Harvesting these links provides useful translation equivalents for many different language pairs, and could provide a basis for further lexical acquisition techniques such as described by Koehn and Knight (2002).

It is expected that the method presented will be particularly useful for under-resourced languages which, in many cases, have an active and vibrant Wikipedia community.

2. Related work

The work presented in this paper is in the same vein as that by Koehn and Knight (2002) in that it focuses on attempting to create a translation lexicon from meagre resources. In their case these meagre resources were unrelated monolingual corpora. They cover a number of methods, some of which were based on linguistic knowledge, and others on statistics.

[1]As used on Wikipedia, a bot (short for robot) is a software program that makes automated changes to the Wikipedia.

Adafre and de Rijke (2006) describe an experiment in finding similar sentences between different language versions of Wikipedia and note that lexicons induced from Wikipedia titles are generally of high quality and there is "rarely conceptual mismatch" between pages linked by `interwiki` links. They propose two approaches, one using a machine translation system and the other using the hyperlinks between documents. Their second approach of working with the hyperlinks within a document is more general and involved than the method we propose here. They do not give any quantitative evaluation of the lexicon created, which includes both common nouns and proper nouns.

Wikipedia has also been used as a semantic resource in the vein of WordNet (Zesch et al., 2007a; Zesch et al., 2007b), and as a monolingual resource in developing systems for named entity and word sense disambiguation (Bunescu and Pașca, 2006; Cucerzan, 2007; Mihalcea, 2007).

3. Method

In this section we shall give a quick overview of the experiment and describe the algorithm used therein.

Our method, as described below, requires a monolingual word list of one of the languages in a translation pair.

Starting from a word list in the better sourced language of the pair is the logical and the recommended practise. In our experiment we had English in all of the language pairs and therefore used an English word list as seed for our method. This word list was extracted from the English–Catalan translation pair of Apertium,[2] an open-source, shallow transfer machine translation system (Armentano-Oller et al., 2005). The motivation behind using this specific wordlist was that the lexicons produced could be immediately useful in Apertium translation pairs. The word list[3] consisted of 11,393 lemmas,[4] all nouns, and was biased slightly towards technical and scientific terminology. The reason for choosing a list made up only of nouns was because Wikipedia titles are almost exclusively made up of nouns and proper nouns. The first ten words are shown below:

```
abandonment
abbey
abbot
abbreviation
abdomen
abduction
aberration
ability
abnormality
abolitionism
```

The total number of articles in the English Wikipedia which matched the entries in the word list was 10,024; this number represents the upper bound on the number of possible translation pairs.

The languages for which translations were attempted to be found were Macedonian (`mk`), Afrikaans (`af`), Iranian Persian (`fa`) and Swedish (`sv`). These choices were motivated by the availability of native speakers to evaluate the results, and the desire to cover a variety of language groups and Wikipedia sizes.

The bilingual word pair extraction algorithm, presented in pseudo-code in Figure 1, is very simple and computationally inexpensive.

```
EXTRACT-WORD-PAIRS()
1  for each w in Word-List
2  do
3     a ← RETRIEVE-PAGE( SourceWikipedia , w);
4     ℓ ← EXTRACT-LINKS(a);
5     for each t in Target-Languages
6     do if t in ℓ
7           then ADD-PAIR(w, ℓ[t]);
```

Figure 1: Description of the algorithm used.

In the algorithm (Figure 1) we iterate over both the word list and list of target languages, represent the extraction of the `interwiki` links with the function EXTRACT-LINKS and the target word/phrase of the `interwiki` link as the array ℓ.

Certain titles are ambiguous (can be associated with more than one topic) and are linked to a page that contain no content and only refers to other Wikipedia articles with which the user can resolve the conflict (Wikipedia, 2008a). In this case the title of these so-called disambiguation pages were taken as the translation. Information within parentheses of titles were uniformly removed from all page titles.

4. Results

The results for precision of this method are presented in Table 1. Also given is the *total* number of articles in the Wikipedia in question (on 9 February 2008), the total number of `interwiki` links retreived and the number of "correct" translations.

Table 1: Results for the language pairs

	Total	Links	Correct	Precision
`af`	9,183	444	354	79%
`mk`	14,887	779	631	81%
`fa`	32,194	1,605	1,487	92%
`sv`	273,291	4,913	3,428	69%
`en`	2,299,336	10,024	-	-

Precision was calculated by dividing the number of correct translations by the total number of possible translations retrieved. A correct translation was counted as an exact lemma-for-lemma translation and was judged by a native speaker. Note that the number of links retrieved, from the English word list of 10,024 entries, is rather low. The scientific and technical nature of the word list could be the cause hereof as more popular topics are added quicker and

[2] Available from `http://www.apertium.org/`

[3] The word list is under the same license as the linguistic package `apertium-en-ca`, and can be retrieved from `http://xixona.dlsi.ua.es/~fran/en-nouns.txt`.

[4] The lemma (or citation form, base form, head word, etc.) is the canonical form of a word, as is typically found in printed dictionaries.

revised more often. As one would expect, the number of pages in the target language's Wikipedia also greatly affects the number of links retrieved.

5. Analysis

The word lists were given to native speakers to check. A positive result is when the translation is judged as correct by a native speaker. That is when the word is in the right form, has the right sense and is in the appropriate register. If a word can have many possible translations, it is considered enough that it be among them, not necessarily being the most general or frequent. As all Wikipedia article titles are in uppercase, case distinctions were ignored. A rough typology for a negative results with examples can be found below:

1. Right sense, wrong surface form – *vandal* translated as *vandale* (vandals). This kind of error occurred when the lexical form of the word in the source language did not match the translation. For example a singular noun being translated as a plural noun.

2. Right sense, wrong register – *nephrolithiasis* translated as *njursten* (kidney stone). This is normally caused by a more scientific term or more specific term being a redirect to a more general article. The English Wikipedia guidelines recommend that the most common name, not the most correct name be used for the title of an article (Wikipedia, 2008c). This problem was also seen in the translation of acronyms, where the acronym typically redirects to the spelt-out form.

3. Wrong sense, right domain – *sociolinguist* translated as *sosiolinguistiek* (sociolinguistics). This type is also generally caused by redirects. Articles on professions, sub-fields, etc. are often redirected to a general article dealing with the whole field. This also occurs with derivations as shown above. It is worth noting that these are by no means regular, for example *bureaucrat* has its own article, while *bureaucratisation* redirects to *bureaucracy*. On the other hand, *colonist* redirects to *colony*, while *colonisation* has its own article.

4. Wrong sense, wrong domain – *solidarity* translated as *Solidarność*.[5] The fourth type of error occurred when an incorrect `interwiki` link was in place. These are caused either by badly configured bots, or human error. Examples of this kind of error are a proper name linked in the place of a common name.

Borderline situations also exist, for example, the translation of *amount* into Macedonian as *kvantitet* (literally 'quantity'). In this example, the translation found is not an exact translation, but refers to a similar and closely related word. These were marked as correct or incorrect translations at the discretion of the native speakers.
These errors were generally found to exist at approximately the same frequency, with none particularly more frequent than the other. No full quantitative analysis was done.

[5] A proper name referring to a trade union, later political party in Poland.

The increase over time in articles and `interwiki` links continue to gradually improve recall (the number of correct translations retrieved from a given word list). Therefore recall will be improved as the number of articles in each Wikipedia grows, along with the number of links between articles. Several techniques could improve precision:

- Double-check each pair – Ensuring that a retrieved link points back to the same source. The equivalent of cross-referencing in a paper dictionary. That is the `interwiki` links of the *target* page are checked for a link to the *source* page.

- Avoid following redirects – This would increase precision at the expense of reducing recall. Often differing orthographic conventions are linked through redirection, and if these links were not followed, the pages would not be retrieved.

- Analyse all links – A more complex strategy might involve retrieving the set of all the `interwiki` links from all the pages linked from the page in the *source* language, and choosing the most frequently linked translation in the *target language*. This is similar to what is done by Adafre and de Rijke (2006).

6. Discussion

We have presented a simple, computationally inexpensive and fast means of automatically obtaining bilingual word lists.
The accuracy of this method compares favourably with those of Koehn and Knight (2002), the lowest accuracy we achieved was 69% compared to the 39% accuracy they obtained in their experiment. But their method operates on unrelated, monolingual corpora and could potentially produce more word pairs.
Extracting word pairs from Wikipedia could prove useful for under-resourced languages, and for bootstrapping more complex induction techniques.
Further work would generally focus on improving the precision of results, although another avenue might be to work with trying to use additional information to provide sense disambiguation for the word pairs. Similar work has been some by Sammer and Soderland (2007), who use bilingual word lists and monolingual corpora to construct a sense disambiguated lexicon. Along with the `interwiki` links, Wikipedia articles are generally members of categories, which could be used for this task. Further disambiguation information comes from the page titles themselves, where there is more than one concept represented by a title, often they are disambiguated by means of a term in parentheses. These terms can be almost anything, indication of hierarchy (in the case of place names), of domain (in the case of nouns), or profession (in the case of people), etc.
Another possible use might be for automatically creating directories for named entities, containing places or people. Wikipedia has large numbers of articles on these topics and often, as they are quite formulaic, they are translated into quite a large number of languages. This strategy has been used in the expansion of dictionary entries for the Occitan – Catalan language pair in the Apertium machine translation

system to improve the coverage of place names. Indeed in further work it might be interesting to compare the accuracy of retrieval of translations of proper nouns to those of common nouns.

The wide range of the precision found, 69–92% would be another avenue for further investigation. Increasing the number of human evaluators of the output would likely provide a more accurate benchmark of translation quality.

A possible caveat with using Wikipedia in this manner is the licensing of the articles. The content of Wikipedia is uniformly released under the GNU Free Documentation Licence (GFDL),[6] which is incompatible with the GNU General Public License (GPL),[7] a licence under which much open-source software, including Apertium, is released. There has been an ongoing discussion of this problem in the Wikipedia mailing lists, however the most authoritative response comes from Mike Godwin, general counsel to the Wikimedia Foundation.[8] He argues that these, "... links and word pairs, standing alone, do not qualify as copyrightable, and thus fall outside the GFDL" (personal correspondence).

Acknowledgements

Many thanks to the people who evaluated the output of the process, Cenny Wenner, Slobodan Jakovski and Soroush Mesry.

7. References

Adafre, S. F. and de Rijke, M. 2006. Finding similar sentences across multiple languages in wikipedia. In *EACL 2006 Workshop on New Text–Wikis and Blogs and Other Dynamic Text Sources*, March.

Armentano-Oller, C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Bonev, B., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., and Sánchez-Martínez, F. 2005. An open-source shallow-transfer machine translation toolbox: consequences of its release and availability. In *OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X*, pages 23–30, September.

Bunescu, R. and Pașca, M. 2006. Using encyclopedic knowledge for named entity disambiguation. *Proceedings of EACL*, pages 9–16.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on wikipedia data. *The EMNLP-CoNLL Joint Conference. Prague*.

Koehn, P. and Knight, K. 2002. Learning a translation lexicon from monolingual corpora. *Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*, 34.

Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation. *Proceedings of NAACL HLT 2007*, pages 196–203.

Sammer, M. and Soderland, S. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. *Proceedings of Machine Translation Summit XI*.

Wikipedia. 2008a. Disambiguation — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Wikipedia. 2008b. List of wikipedias — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Wikipedia. 2008c. Naming conventions (common names) — Wikipedia, the free encyclopedia. [Online; accessed 9 February 2008].

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007a. Analyzing and Accessing Wikipedia as a Lexical Semantic Resource. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications*, pages 197–205. Gunter Narr, Tübingen, Tuebingen, Germany.

Zesch, T., Gurevych, I., and Mühlhäuser, M. 2007b. Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, pages 205–208.

[6] Available at `http://www.gnu.org/licenses/fdl.html`.

[7] Available at `http://www.gnu.org/licenses/gpl.html`.

[8] The Wikimedia Foundation operates Wikipedia and related sites.