# Exploiting a Multilingual Web-based Encyclopedia for Bilingual Terminology Extraction

Fatiha Sadat
University of Quebec in Montreal, Computer Science department,
201 President Kennedy avenue, Montreal, QC, Canada

sadat.fatiha@uqam.ca

**Abstract.** Multilingual linguistic resources are usually constructed from parallel corpora, but since these corpora are available only for selected text domains and language pairs, the potential of other resources is being explored as well. This article seeks to explore and to exploit the idea of using multilingual web-based encyclopedias such as Wikipedia as comparable corpora for bilingual terminology extraction. We propose an approach to extract terms and their translations from different types of Wikipedia link information and data. The next step will be using a linguistic-based information to re-rank and filter the extracted term candidates in the target language. Preliminary evaluations using the combined statistics-based and linguistic-based approaches were applied on different pairs of languages including Japanese, French and English. These evaluations showed a real open improvement and a good quality of the extracted term candidates for building or enriching multilingual ontology, dictionaries or feeding a cross-language information retrieval system with the related expansion terms of the source query.

**Keywords:** Bilingual terminology, comparable corpora, Wikipedia, multilingual linguistic tool.

## 1 Introduction

In recent years two types of multilingual corpora have been an object of studies and research related to natural language processing and information retrieval: parallel corpora and comparable corpora. The parallel corpora are made up of original texts and their translations (Morin et al., 2004 ; Véronis, 2000). This allows texts to be aligned and used in applications such as computer-aided translator training and machine translation systems. This method could be expensive for any pair of languages or even not applicable for some languages, which are characterized by few amounts of Web pages on the Web. On the other hand, non-aligned comparable corpora, more abundant and accessible resources than parallel corpora, have been given a special interest in bilingual terminology acquisition and lexical resources enrichment (Dejean et al., 2002; Fung, 2000; Gœuriot et al., 2009a; Gœuriot et al., 2009b; Morin et al., 2006; Nakagawa et al., 2000; Rapp, 1999; Sadat et al., 2003; Sadat et al., 2004). Comparable corpora are defined as collections of texts from pairs or multiples of languages, which can be contrasted because of their common features in the topic, the domain, the authors, the time period, etc. Comparable corpora could be collected from downloading electronic copies of newspapers and articles, on the WWW for any specified domain.

Among the advantages of comparable corpora; their availability, consistency and utility for research on Natural Language Processing (NLP). In another hand, recent publications on bilingual terminology extraction from comparable corpora have shown promising results although most used comparable corpora are domain-specific, which causes limitations on the usage diversity, the domain and the quality of terminology.

This paper intends to bring solutions to the problem of lexical coverage of existing linguistic resources such as multilingual ontologies and dictionaries, but also to the improvement of the performance of Cross-Language Information Retrieval. The main contribution of the current

study is an automatic acquisition of bilingual terminology from Wikipedia[1] articles in order to construct a bilingual ontology or enhance the coverage of existing ontolgies.
The remainder of the present paper is organized as follows: Section 2 presents an overview of Wikipedia. Section 3 presents the different steps for the acquisition of bilingual terminology using a two-stage corpus-based translation model. Experiments and evaluations are related in Section 4. Section 5 concludes the present paper.

## 2 An overview of Wikipedia

Wikipedia (*having the pronunciation wikipe'dja or vikipe'dja*) is an online multilingual encyclopedia based on the Internet, universal, multilingual and working on the concepts of a wiki, i.e. a web site with freely updatable web pages from all or a part of visitors of that site.
Wikipedia offers a gigantesque repository of multilingual data to exploit automatically for different aims in NLP. Different search engines such as *Google*[2] or *Yahoo*[3] or Wikipedia's can be used for the implementation of the approach to extract bilingual terminology from comparable corpora and its related evaluations.
Wikipedia offers a neutral content that can be verified and updated freely by any editor. The edition of collaborative documents can be monolingual or multilingual. Actually, the French version of Wikipedia (francophone[4]) has more than 943 399 articles and more than 5 000 actives contributors[5].
This considered linguistic resource can be used as parallel or comparable corpora: it can be considered as gigantesque lexical resource, available freely for all users, for many domains and diverse languages. However, its exploitation in NLP research is recent, not completely pertinent and still requires theoretical ideas and practice on its statute, characteristics and limits (Adafre et al., 2006; Schönhofen et al., 2007; Erdmann et al., 2008a; Erdmann et al., 2008b; Erdmann et al., 2009; Adar et al., 2009; Mohammadi et QasemAgharee, 2009 ; Yu et Tsujii, 2009a; Yu et Tsujii, 2009b).
The aim of the current study is the acquisition of bilingual or multilingual terminology from Wikipedia articles, which is automatic and language independent. The evaluation of our ideas and approach is done on different pairs of languages including French, English and Japanese.
According to figure 1, the number of Wikipedia articles for most of European languages, has achieved a limit that allows this resource to be used in NLP research and more specifically in multilingual information extraction and retrieval. Although, the advance of this resource content, most of studies were concentrated on the monolingual aspect (Voss 2005). We are interested in the multilingual aspect of Wikipedia in order to extract the pertinent terminology for the development of a multilingual ontology or dictionary.

## 3 The approach of bilingual terminology extraction

The process to extract bilingual terminology from Wikipedia documents is described as follows: (1) construction of comparable corpora; (2) translation using a statistical approach; (3) combination to linguistic information in order to filter and re-rank the extracted terminology as described in Sadat et al. (2003, 2004).

[1] http://www.wikipedia.org
[2] http://www.google.com
[3] http://www.yahoo.com
[4] http://fr.wikipedia.org
[5] Information of April 30th 2010 at 9:34 am

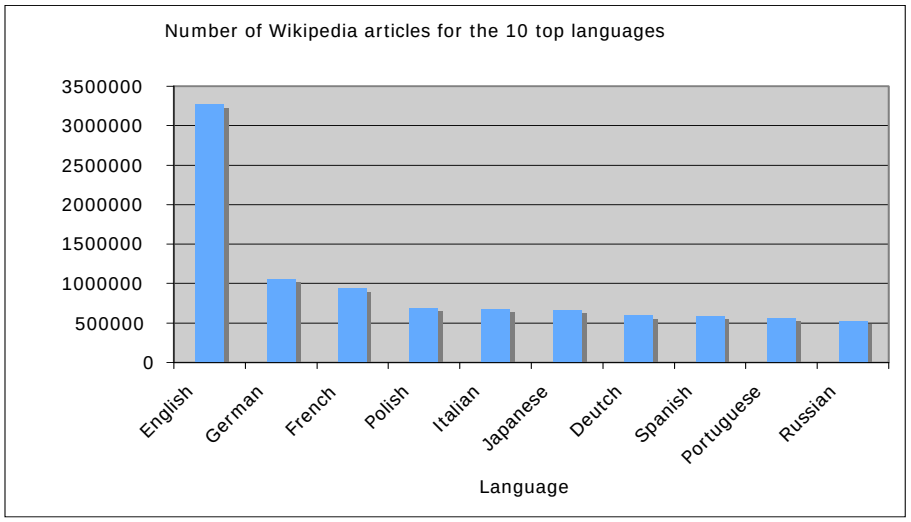


**Figure 1:** Number of Wikipedia articles for the 10 most used languages

First, we consider a preliminary query *Q* of *n* words in a source language *S* to input in Wikipedia search engine. The resulted document is used as a first document for the corpus in the source language *S*. The usage of the inter-language link in the target language *T* for this document will lead to a corpus in a target language *T*.
Following this first step and exploiting the links in the same document as well as the inter-language links, comparable corpora are built for the query *Q*.
In this study, we use the term *deep* in the same document to define the number of times links in the same language are used in our corpora. Example, a first document $corp_1$ is described by $deep_0$; using the *i* links that are included in this first document will lead to $deep_1$ and to an extended corpus $corp_2$, using the *j* links that are included in the $corp_2$ will lead to $deep_2$ and to an extended corpus $corp_3$. This step can be terminated at $deep_m$ to result in a comparable corpus $corp_{m+1}$.
The exploitation of all links related to the set [$deep_0$ … $deep_m$ ] in the source language will lead to a corpus in this language. In another hand, a parallel exploitation of inter-language links in the target language for the same set [$deep_0$ … $deep_m$ ] will lead to a corpus in the target language that is comparable to the one in the source language.
This approach can be used to build a multilingual corpus in several languages according to the availability of documents in all those languages in Wikipedia sites.
Second, the statistical phase will realize the alignment between terms of the source language and those in the target language. Considering the constructed comparable corpora from Wikipedia articles, we apply the following steps to extract the bilingual terminology:

1. *Extraction of terms from source and target languages documents:* In this step, terms with the following part of speech tags are extracted: *noun, verb, adverb, adjective*.
2. *Construction of context vectors in both languages:* For each term *w*, a context vector is constructed considering all terms that co-occur with the term *w* in a specified window size of one phrase. The mutual information (Dunning, 1993) is used as a co-occurrence tendency measure.
3. *Translation of the context vector content in the source language to the target language*: Context vectors of words in the source language are translated into the target language using the Wikipedia resource as translator. This step requires using the interlink information of Wikipedia for word translation. If needed, the *Wikitionaire*[6] is used to overcome the limitations of Wikipedia and to deal with out-of-vocabulary words. In the

[6] http://fr.wiktionary.org/

current study, we are interested by exploiting specifically Wikipedia as a multilingual and lexical resource, although it is possible to use a bilingual dictionary or a freely available machine translation to overcome the limitations of the translation.

4. *Construction of similarity vectors*: Context vectors (original and translated) of words in both languages are compared using the *cosine metrics*. Other measures such as the *Jaccard distance* or the *Dice coefficient* can be considered.

The third step consists on a *linguistics-based pruning approach*. Terms and their translations that are morphologically close enough, i.e., close or similar POS tags, are filtered and retained. We restricted the pruning technique to nouns, verbs, adjectives and adverbs, although other POS tags could be treated in similar way.

Finally, the generated translation alternatives are sorted in decreasing order by similarity values. Rank counts are assigned in increasing order, starting at 1 for the first sorted list item. A fixed number of top-ranked translation alternatives are selected and misleading candidates are discarded.

In this proposed approach, all monolingual links in a document are used to extract terms and concepts in the related language. In another hand, links involving two or several languages are used to retrieve terms across those languages.

## 4 Evaluations

Our preliminary evaluations using the proposed strategy were based on different pairs of languages including French, English and Japanese. Different sizes of the Wikipedia corpus were used and referenced here by the term *deep*.

Table 1 shows the size of the bilingual corpora according to the exploitation of same-language links and inter-language links.

Tables 2,3 and 4 show the results of the obtained bilingual terminology according to different sizes of the corpora for the French-English, Japanese-French and Japanese-English pairs of languages, respectively.

Note that we used a first query including the terms « *infection hospital illness tuberculosis* » which is a part of NTCIR-7[7] test collection in CLIR, in the three languages, i.e. French, Japanese and English.

Table 5 shows an example for the extracted bilingual terminology in English for the source term « *santé* » in French (which means *health* in English) with $deep_3$.

The obtained terminology is very useful for building a bilingual ontology (or multilingual). The extracted terms have a certain semantic relationship with the source term and the resulted documents in the source and target languages can be exploited in order to define the semantic relations and thus build a multilingual ontology.

**Table 1:** Sizes of Wikipedia corpora according to different links exploitation

| Deep | Number of tokens/articles in French | Number of tokens/articles in English | Number of tokens/articles in Japanese |
|---|---|---|---|
| 0 | 388 / 4 | 510 / 4 | 57 / 4 |
| 1 | 4 511 / 61 | 5 633 / 51 | 185 / 10 |
| 3 | 161 967 / 2 634 | 205 023 / 2 121 | 10 964 / 266 |
| 7 | 533 931 / 9 077 | 657 035 / 7 110 | 45 378 / 977 |

[7] http://aclia.lti.cs.cmu.edu/ntcir8

**Table 2:** Examples of the extracted bilingual terminology (French-English) according to different sizes of the Wikipedia corpus

| Deep | Source term (fr.) | Number of candidates (eng.) | Ideal translation (eng.) | Rank |
|---|---|---|---|---|
| 0 | organisation | 14 | Organization | 1 |
| | organisation | 14 | Institution | 4 |
| | organisation | 14 | compagny | 8 |
| | organisme | 105 | organism | 4 |
| | Maladie | 101 | Disease | 14 |
| | Santé | 89 | Health | 1 |
| | hôpital | 19 | Hospital | 3 |
| 1 | admission | 90 | admission | 1 |
| | algue | 88 | Algae | 1 |
| | thérapie | 52 | therapy | 1 |
| | animal | 369 | animal | 2 |
| | assistance | 85 | support | 3 |
| | blessure | 269 | injury | 3 |
| | épidemiologie | 186 | epidemiology | 5 |
| 3 | abeille | 443 | bee | 1 |
| | narcotique | 1656 | narcotic | 1 |
| | assurance | 289 | insurance | 2 |
| | chimie | 2044 | chemistry | 3 |
| | silicone | 132 | silicone | 3 |
| | médecine | 416 | medicine | 4 |
| | réanimation | 1004 | resuscitation | 5 |
| | taxonomie | 1841 | taxonomy | 7 |

**Table 3:** Examples of the extracted bilingual terminology (Japanese-French) according to different sizes of the Wikipedia corpus

| Deep | Source term (jap.) | Number of candidates (fr.) | Ideal translation (fr.) | Rank |
|---|---|---|---|---|
| 0 | 感染 | 14 | Infection | 1 |
| 1 | イングランド | 16 | Angleterre | 2 |
| | けっか | 6 | Résultat | 1 |
| | 世界 | 14 | Monde | 1 |
| 3 | アレルギー | 236 | Allergie | 2 |
| | セルロース | 233 | Cellulose | 2 |
| | ワクチン | 102 | Vaccin | 1 |

**Table 4:** Examples of the extracted bilingual terminology (Japanese-English) according to different sizes of the Wikipedia corpus

| Deep | Source term (jap.) | Number of candidates (eng.) | Ideal translation (eng.) | Rank |
|---|---|---|---|---|
| 0 | 細菌 | 17 | Bacteria | 2 |
| 1 | 感染 | 36 | Infection | 2 |
| | ドイツ | 4 | Germany | 1 |
| 3 | ワクチン | 88 | Vaccine | 2 |
| | ヒト | 472 | Human | 1 |
| | 微生物 | 84 | Microorganism | 1 |

**Table 5:** Example of the extracted bilingual terminology in English for the term « *santé* » in French (*deep*$_3$)

| Source term (Fr.) | Translation candidate (Eng.) | Cosinus | Jaccard distance | Dice coefficient |
|---|---|---|---|---|
| Santé | creation | 1,42156726 | 0,01923077 | 0,03773585 |
| Santé | foundation | 1,49393827 | 0,03738318 | 0,07207207 |
| santé | preventive | 1,49393827 | 0,03738318 | 0,07207207 |
| santé | staff | 1,50187875 | 0,03278689 | 0,06349206 |
| santé | health | 1,50243129 | 0,065 | 0,12206573 |
| santé | medicine | 1,50630024 | 0,06410256 | 0,12048193 |
| santé | confusion | 1,50634785 | 0,00943396 | 0,01869159 |
| santé | component | 1,51284513 | 0,02479339 | 0,0483871 |
| santé | charge | 1,51350091 | 0,00862069 | 0,01709402 |
| santé | treatment | 1,52215418 | 0,06024096 | 0,11363636 |
| santé | hospital | 1,54637259 | 0,04494382 | 0,08602151 |
| santé | spécialisé | 1,54749611 | 0,01941748 | 0,03809524 |
| santé | epidemiology | 1,55022665 | 0,02654867 | 0,05172414 |
| santé | patient | 1,55067588 | 0,03351955 | 0,06486486 |
| santé | equipment | 1,55167781 | 0,01801802 | 0,03539823 |
| santé | risk | 1,55209348 | 0,03539823 | 0,06837607 |
| santé | approach | 1,55211298 | 0,03603604 | 0,06956522 |

## 5 Conclusion

In this paper, we investigated the approach of extracting bilingual terminology from Wikipedia documents as comparable corpora in order to enrich and/or construct bilingual ontologies. We proposed a simple and adaptable approach to any language and showed preliminary evaluations for three pairs of languages including French, Japanese and English. This proposed approach showed promising results for this first study.

Among the drawbacks of the proposed approach is the introduction of many noisy terms or wrongly translations; however, most of those terms could be considered as efficient for the definition of semantic relationships in order to enrich an ontology in bilingual or multilingual format.

Further extensions include more evaluations to determine the precision and quality of translation as well as the performance of the whole system. Also, we are interested by the

decomposition of the constructed large corpora using Wikipedia documents into comparable pieces or paraphrases, instead of taking the whole corpus as a single piece. Last, our main objective is the construction of a multilingual ontology and a study of several languages including those with complex morphology, such as Arabic.

## References

Adafre S. F., De Rijke, M. (2006). Finding similar sentences across multiple languages in Wikipedia. In Proceedings of the EACL Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources.

Adar E., Skinner M., Weld D. S. (2009). Information arbitrage across multi-lingual Wikipedia, Proceedings of the Second ACM International Conference on Web Search and Data Mining, February 09-12, 2009, Barcelona, Spain.

Dejean, H., Gaussier, E., Sadat, F. (2002). An Approach based on Multilingual Thesauri and Model Combination for Bilingual Lexicon Extraction. In Proceedings of COLING'02, Taiwan.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational linguistics 19(1): 61-74.

Erdmann M., Nakayama K., Hara T., Nishio S. (2008a). An approach for extracting bilingual terminology from Wikipedia. In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA).

Erdmann M., Nakayama K., Hara T., Nishio S. (2008b). Extraction of bilingual terminology from a multilingual Web-based encyclopedia. J. Inform. Process.

Erdmann M., Nakayama K., Hara T., Nishio S. (2009). Improving the extraction of bilingual terminology from Wikipedia. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), Volume 5 , Issue 4 (October 2009). 2009

Fung, P. (2000). A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In Jean Véronis, Ed. Parallel Text Processing.

Gœuriot, L., Daille, B., and Morin, E. Compilation of specialized comparable corpus in French and Japanese. Proceedings, *ACL-IJCNLP workshop "Building and Using Comparable Corpora" (BUCC 2009)*, august 2009, Singapore. (2009).

Gœuriot, L., Morin, E., and Daille, B. Reconnaissance de critères de comparabilité dans un corpus multilingue spécialisé. Actes, *Sixième édition de la Conférence en Recherche d'Information et Applications (CORIA 2009)*. (2009) .

Kun Y., Tsujii J. (2009). Bilingual Dictionary Extraction from Wikipedia. (2009a). in Proceedings of MT Summit XII proceedings 2009.

Kun Y., Junichi T. (2009b). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. in Proceedings of NAACL HLT 2009: Short Papers, pages 121–124, Boulder, Colorado, June 2009.

Mohammadi M., QasemAgharee N. (2009). In proceedings of NIPS Workshop, Grammar Induction, Representation of Language and Language Learning. December 2009, Whistler, Canada.

Morin, E., Daille, B. Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé. *Traitement Automatique des Langues (TAL)*, Lavoisier, 45(3), 103–122. (2004)

Morin, E., and Daille, B. (2006). Comparabilité de corpus et fouille terminologique multilingue. Traitement Automatique des Langues (TAL), 47(1):113-136, 2006.

Nakagawa, H. Disambiguation of Lexical Translations Based on Bilingual Comparable Corpora. In *Proceedings of LREC2000, Workshop of Terminology Resources and Computation WTRC2000*, pp 33-38. (2000)

Peters, C., Picchi, E. Capturing the Comparable: A System for Querying Comparable Text Corpora. In *Proceedings of the Third International Conference on Statistical Analysis of Textual Data*, pp 255-262. (1995)

Rapp, R. (1999). Automatic Identification of Word Translations from Unrelated English and German Corpora. In Proceedings of European Chapter of the Association for Computa-tional Linguistics EACL.

Sadat, F., Yoshikawa, M., Uemura, S. (2003). Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach. In Proceedings of EACL'203, workshop on Information retrieval with Asian languages - Volume 11, Sapporo, Japan. Pages: 57–64.

Sadat, F. (2004). Knowledge Acquisition from Collections of News Articles to Cross-language Information Retrieval. In Proceedings of RIAO 2004 conference, Avignon, France, pp. 504-513.

Véronis, J. (2000). Parallel Text Processing: Alignment and Use of Translation Corpora. Dordrecht, Kluwer Academic Publishers Ed 2000.

Voss J. (2005). Measuring Wikipedia. In Proceedings of 10th International Conference of the International Society for Scientometrics and Informetrics.