Exploiting Wikipedia as a Knowledge Base for the Extraction of Linguistic Resources: Application on Arabic-French Comparable Corpora and Bilingual Lexicons

Rahma Sellami
ANLP Research Group
Laboratoire MIRACL
University of Sfax, Tunisia
rahma.sellami@gmail.com

Fatiha Sadat
UQAM, 201 av. President Kennedy,
Montreal, QC, H3X 2Y3, Canada
sadat.fatiha@uqam.ca

Lamia Hadrich Belguith
ANLP Research Group
MIRACL Laboratory
University of Sfax, Tunisia
l.belguith@fsegs.rnu.tn

Abstract

We present simple and effective methods for extracting comparable corpora and bilingual lexicons from Wikipedia. We shall exploit the large scale and the structure of Wikipedia articles to extract two resources that will be very useful for natural language applications.

We build a comparable corpus from Wikipedia using categories as topic restrictions and we extract bilingual lexicons from inter-language links aligned with statistical method or a combined statistical and linguistic method.

1 Introduction

Multilingual linguistic resources are usually constructed from parallel corpora. Unfortunately, parallel texts are scarce resources: limited in size, language coverage, and language register. There are relatively few language pairs for which parallel corpora of reasonable sizes are available.

The lack of these corpora has prompted researchers to exploit other multilingual resources such as comparable corpora. Comparable corpora are "sets of texts in different languages that are not translations of each other" (Bowker and Pearson, 2002), but contains texts from the same domain.

Comparable corpora have several obvious advantages over parallel corpora. They are available on the Web in large quantities for many languages and domains and many texts with similar content are produced every day (e.g. multilingual news feeds) (Skadiņa et al, 2010), but they are not organized.

Also, bilingual lexicons are the key component of all cross-lingual NLP applications such as machine translation (Och and Ney, 2003) and cross-language information retrieval (Grefenstette, 1998).

Parallel texts – as the most important resource in statistical machine translation (SMT) – appear to be limited in quantity, genre and language coverage. Providing more comparable corpora essentially boosts the coverage and the quality of machine translation system, especially for less-covered languages and domains.

In this paper we describe the extraction process of large comparable corpora and bilingual lexicons for Arabic and French language from a multilingual web-based encyclopedia, Wikipedia.

We propose to build bilingual resources as follows: first comparable corpora from Wikipedia using categories and languages as restrictions; next two bilingual lexicons extracted from titles of articles that are related by inter-language links and aligned by a statistical based method and a combined statistical and linguistic-based method.

The best extracted lexicon will be used to improve the mining of different levels of parallelism from our comparable corpora.

The content of this paper is summarized as follows: Section 2 describes some characteristics of Wikipedia that makes it a source of multilingual resources extraction. Section 3 presents a brief overview of previous works on comparable corpora and bilingual lexicon extraction from Wikipedia. In sections 4 and 5, we present and evaluate our work of mining Arabic-French comparable corpora and bilingual lexicon from Wikipedia. We conclude the present paper in section 6.

2 Characteristics of Wikipedia

In the following sub-section, we shall describe some of the interesting characteristics of Wikipedia that make the encyclopedia an invaluable resource for knowledge mining.

Wikipedia is an online encyclopedia under the non-profit Wikimedia Foundation. Unlike ordinary encyclopedias, the Wikipedia project is based on the wiki concept (Leuf and Cunningham, 2001), thus anyone can contribute by creating, editing or improving the articles.

2.1 Wikipedia Coverage

Wikipedia currently (2012) contains more than 22 million articles among which 1 259 482 are written in French and 179 291 are written in Arabic language[1]. These articles cover different categories such as arts, geography, history, society, science and technology. Wikipedia articles cover many domain-specific concepts as well as named entities (i.e. proper nouns such as names of persons), including even latest topics since Wikipedia is being updated all the time.

2.2 Wikipedia Link Structure

- Inter-language Links

An inter-language link in Wikipedia is a link between two articles in different languages. An article has usually one inter-language link for each language.

Inter-language links are created using the syntax *[[language code:article title]]*. The *language code* identifies the language in which the target article is written and *Article title* is the title of the target page (e.g. [[fr:Lac Tchad]]). Since the titles of all Wikipedia articles in one language are unique, that information is sufficient to identify the target page unambiguously.

- Redirect Pages

A redirect is a page, which has no content itself, but sends the reader to another article, section of an article or page, usually from an alternative title. A redirect page can be created by writing the text *#REDIRECT [[article title]]* at the top of the article where *article title* denotes the name of the target page.

Redirect pages are used in particular for Adjectives/Adverbs point to noun forms (e.g. *Treasonous* redirects to *Treason*), Abbreviations (e.g., *DSM-IV* redirects to *Diagnostic and Statistical Manual of Mental Disorders*), Alternative spellings or punctuation (e.g. *Al-Jazeera* redirects to *Al Jazeera*), etc.

- Link Texts

This is a link to another page in Wikipedia. The link text can correspond to the title of the target article (the syntax will be: *[[article title]]*), or differ from the title of the target article (with the following syntax: *[[article title | link text]]*).

As a rich and free resource, Wikipedia has been successfully used as an external resource in many natural language processing tasks (Buscaldi and Rosso, 2006; Mihalcea, 2007; Nakayama et al., 2007).

3 State of the Art

In accordance with fast growth of Wikipedia, many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: in this paper, our main concern is the use of Wikipedia as a source of comparable corpora and bilingual lexicon extraction.

Li et al. (2010) consider Wikipedia as a comparable corpus, they align articles pairs based on inter-language links for the extraction of parallel sentences. Patry and Langlais (2011) also concentrate on documents pairs that are linked across language for extracting parallel documents. However, Smith et al. (2010) and Mohammadi and QasemAghaee (2010) use inter-language link to

[1] http://meta.wikimedia.org/wiki/List_of_Wikipedias

identify aligned comparable Wikipedia documents. Sadat (2010) proposes an approach to build comparable corpora from Wikipedia encyclopedia. First, the author considers a preliminary query Q in a source language to input in Wikipedia search engine. The resulting document is used as a first document for the corpus in the source language. The usage of the inter-language link in the target language for this document leads to a corpus in a target language. 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.

Otero and Lopez (2010) propose an automatic method to build comparable corpora (CorpusPedia) from Wikipedia using Categories as topic restrictions. Given two languages and a particular topic, their strategy builds a corpus with texts in the two selected languages, whose content is focused on the selected topic. Again, Otero and Lopez (2011) propose two strategies to build comparable corpora from Wikipedia: The first one (non-aligned corpus) extracts those articles in two languages having in common the same topic. It results in a non-aligned comparable corpus, consisting of texts in two languages. The second strategy (aligned corpus) extracts pairs of bilingual articles related by inter-language links, with the condition that at least one of both contains a required category. It results in a comparable corpus with aligned articles. The input of the two strategies is CorpusPedia developed by Otero and Lopez (2010).

Plamada and Volk (2012) demonstrate the difficulty to use Wikipedia categories for the extraction of domain-specific articles from Wikipedia. They propose an Information Retrial (IR) approach in order to achieve a solution to this task and they identify articles that belong to the Alpine domain based on this approach.

Skadina et al., (2012) developed a technique to find comparable Wikipedia texts based on inter-language link. First, they extract all document pairs connected by inter-language link and share the same topic. Then, they filter out non-comparable articles; they measure the similarity of document pairs by performing cross-lingual sentences alignment.

Several works have a common characteristic: their comparable corpora are composed from articles related by inter-language links that may share or not the same topic. However, our work is based on the definition of comparable corpora, a set of texts that share some criteria without being in mutual translation. We constructed a comparable corpus from articles that share at least one topic, but are not necessary related by any inter-language link.

Other works on the extraction of bilingual lexicons from Wikipedia are described as follows: Adafre and Rijke (2006) created a bilingual dictionary (English-Duch) from Wikipedia in order to help construct a parallel corpus. The authors demonstrated that the bilingual lexicon approach for constructing a parallel corpus is more accurate and efficient than the machine translation based approach. Bouma et al. (2006) extracted bilingual terminology for creating a multilingual question answering system (French-Dutch). In addition, Decklerck et al. (2006) used bilingual terminology for translating ontology labels; they used only inter-language links for bilingual terminology extraction.

What all researches have in common is the fact that they use only inter-language links for extracting bilingual terminology. However, Erdmann et al. (2008) analyze not only the inter-language link of Wikipedia, but also exploit redirects links and link texts to build an English-Japanese dictionary. The authors have shown the contribution of using Wikipedia compared to parallel corpus for the extraction of a bilingual dictionary. This contribution appears especially at the wide coverage of terms.

Sadat and Terrasa (2010) propose an approach for extracting bilingual terminology from Wikipedia. This approach, first, extract pairs of words and translations from different types of information, links and text of Wikipedia, then, use linguistic information to reorder the relevant terms and their translations. More recently, Ivanova (2012) evaluates a bilingual bidirectional English-Russian dictionary created from titles of Wikipedia articles. She explored the inter-language links and redirect pages methods described in (Erdmann et al., 2008) in order to create English-Russian Wiki-dictionary. The author demonstrates that Machine translation experiments with the Wiki-dictionary incorporated

into the training set resulted in the rather small, but statistically significant drop of the quality of translation compared to the experiment without the Wiki-dictionary. However, using the test set collected from Wikipedia articles, the model with incorporated dictionary performed better.

4 Comparable Corpora Extraction

4.1 Extraction Process

In this paragraph, we describe our method for building a comparable corpus from Wikipedia articles. This method extracts those articles in two languages having in common the same topic where the topic is represented by a category and its translation.

The process to extract Arabic-French comparable corpora from Wikipedia is described as follows:

First step consists on downloading French and Arabic Wikipedia database (January / February 2012) from http://download.wikimedia.org.

Second, all Arabic topics that have French translations are extracted from Wikipedia articles. An example is the phrase title in Arabic "تصنيف: لاعبو كرة مضرب ألمان", that leads to a French translation *"joueur allemand de tennis"*, when following the syntax of inter-language links.

Third, for each pair of topics, we extract all Arabic and French articles which have a link text to the selected topic.

Fourth step consists on cleaning the extracted articles by removing Wikipedia markups.

Through these steps, we get a comparable corpus of texts in Arabic and French languages sharing the same topic. The comparable corpus covers all topics that exist in Wikipedia.

We should note that there are articles in Arabic, respectively in French, with no corresponding version in French, respectively Arabic.

4.2 Experiments and Results

We download Arabic and French Wikipedia database (January/February 2012) in XML format from http://download.wikimedia.org.

We extract 20 533 Arabic topics that have translation in French language.

In order to have an idea about the size of our corpus, we present the number of Arabic and French articles for the first ten extracted topics. Table 1 summarizes the quantitative description of generated corpora.

Category	Number of Arabic articles	Number of French articles
بحيرات / Lac 'Lake'	41	9
حروب / Guerre 'War'	51	66
رؤساء مصر / Président d'Égypte 'President of Egypt'	5	5
نازية / Nazisme 'Nazism'	11	52
فلك / Astronomie 'astronomy'	255	47
فلاسفة / Philosophe 'philosopher'	40	5
عناصر كيميائية / Élément chimique'chemical element'	168	165
لغات برمجة / Langage de programmation 'Programming language'	39	260
قارات / Continent 'Continent'	29	12
لغات / Langue 'language'	80	32
Total	**719**	**653**

Table 1. Number of Arabic and French articles for the first ten extracted topics.

The table shows that there are significant differences in term of the size among the Arabic and French language, e.g. 41 Arabic articles are sharing the category "بحيرات/Lac 'Lake'" against only 9 in French. However the difference between Arabic and French is less expected since we extract these topics from an Arabic database to seek French articles that share the same topics.

5 Bilingual Lexicon Extraction

5.1 Extraction Process

We propose to use a simple but effective method for bilingual lexicon extraction; it exploits inter-language links between Wikipedia articles to extract Arabic terms (simple or multi-word) and their translations into French. We then use a statistical approach for aligning words of compound terms. Also, linguistic-based filtering based on the part of speech can be applied in order to keep pertinent translation candidates.

We analyze all inter-language links in Wikipedia to create an Arabic-French lexicon. These links are created by the authors of the articles; we assume that the authors correctly positioned these links. Also, an article in the source language is linked to a single article in the target language. Therefore, possible problems of ambiguity in the extraction of pairs of titles are minimized.

We start by downloading Wikipedia database (January 2012) in XML format and thus extracting about 104 104 (Arabic-French) inter-language links. Each link corresponds to a pair of Arabic-French titles.

Some titles are composed of a simple word, while others are composed of multi words. We performed an alignment step in order to have a lexicon consisting only of simple words.

Before aligning these titles, we proceed to preprocessing them. The preprocessing step consists of removing all Arabic and French stop words.

The step of word alignment presents several challenges. First, the alignments are not necessarily contiguous. Two consecutive words in the source sentence can be aligned with two words arbitrarily distant from the target sentence. This is called distortion. Second, a source language word can be aligned to many words in the target language; that is defined as fertility.

The alignment of words of each title is based on IBM models [1-5] (Brown et al., 1993) in combination with the Hidden Markov Model (Vogel et al., 1996). These standard models have already proven their effectiveness in many researches.

The five IBM models estimate the probability P(fr|ar) and P(ar|fr), for which fr is a French word and ar is an Arabic word. Each model is based on the parameters estimated by the previous model and incorporates new features such as distortion, fertility, etc.

The Hidden Markov Model (HMM usually appointed) (Vogel et al., 1996) is an improvement of IBM2 model. It explicitly models the distance between the alignment of the current word and an alignment of the previous word.

We used the open source toolkit GIZA++ (Och and Ney, 2003) that is an implementation of the original IBM models. Thus, GIZA++ was run in both directions to obtain two GIZA++ alignments (Arabic to French and French to Arabic).

Next, two different methods have been exploited in order to filter the two lexicons. First, a combined statistical method with a grow-diag-final heuristics will keep the intersection of the two alignments and thus add additional alignment points. In total we extracted 224 379 translation pairs.

Second a linguistic-based method will keep translation candidates with corresponding part of speech tags of both Arabic and French alignment and will discard non pertinent translations. Stanford Part-Of-Speech Tagger[2] is used for both languages. In total we extracted 235 938 translation pairs.

5.2 Evaluation

Since the titles of Wikipedia articles are usually nouns, our lexicon does not contain verbs.

We calculated the standard criteria precision to measure the accuracy of our methods for the extraction of Arabic-French lexicon from Wikipedia. Our result is based on the precision measure that calculates how many of the extracted translation candidates are correct, as follows:

$$\text{precision} = \frac{|\text{Extracted correct translations}|}{|\text{All extracted translation candidates}|}$$

It is not trivial to estimate the total number of correct translations for a term. Since it cannot be calculated automatically, we conduct a manual

[2] http://nlp.stanford.edu/software/tagger.shtml

evaluation with a support from an expert. We calculate the precisions of our two lexicons based on the candidate translations of 50 words and we compare it to the precision of the online LAROUSSE[3] dictionary.

Table 2 summarizes a comparative description of generated lexicons.
The combined statistical and linguistics based methods that is enhanced with part of speech (POS) filtering achieved a better precision than the stand-alone statistical method.

Statistical method + POS	
candidates	precision
189	80.15%
Statistical method	
candidates	precision
237	76.02%
LAROUSSE	
candidates	precision
66	95,45%

Table 2. Evaluation based on the candidates translations of 50 French words.

The coverage value represent the number of candidates translations, it is 224 379 for the lexicon based on the statistical method and 235 938 for the lexicon based on a combined statistical and linguistics based method using the part of speech filtering.

Mistranslations of our Arabic-French lexicons are mainly due to the fact that some articles' titles are introduced in language other than Arabic (e.g. cv / cv), mostly in English and some translations candidates are transliteration of Arabic word (e.g. Intifada / انتفاضة). Also, we detected alignment errors (e.g. نفسية/ diagnostic). Other errors are due to the fact that pairs of titles are not accurate translations but refer mainly to the same concept (e.g. Christmas / عيد).

6 Conclusion and Future Work

The semi-structured information underlying Wikipedia turns out to be very useful to build multilingual resources such as comparable corpora, parallel corpora, multilingual lexicons and ontologies.

In this paper, we presented our preliminary work on mining Wikipedia for the extraction of comparable corpora and bilingual lexicons. Our major goal is to exploit the multilingual aspect of Wikipedia for Statistical Machine Translation.

On the one hand, we exploit the classification of articles with categories corresponding to topics or genders to extract an Arabic-French comparable corpus. On the other hand, we exploit the network of inter-language links to create an Arabic-French bilingual lexicon.

Unlike previous works that exploit inter-language link to construct comparable corpora, we have tried to build our comparable corpus by selecting Arabic and French articles that share at least one topic. This strategy improves the coverage of comparable corpora. Indeed, even articles that share the same topic despite not related by any inter-language link may contain parallel sentences.

Also, the proposed methods of bilingual lexicon extraction are effective despite its simplicity. We extract Arabic and French articles' titles based on inter-language links between Wikipedia articles. We align words of these titles based on statistical method first; then based on a combined statistical and linguistics based method using the part of speech filtering.

We have reached encouraging levels of precision and coverage, mainly for the second method. These levels exceed respectively 90% and 235 938 pairs of translations for the combined statistical and linguistics-based method using the part of speech filtering.

Finally, in future work, we will define an evaluation protocol to measure the degree of comparability between texts of our comparable corpus. For this purpose, we will make use of techniques described in (Otero and Lopez, 2007) which take advantage of the translation equivalents inserted in Wikipedia by means of inter-language links. We also plan to expand the coverage of our lexicon by exploiting other links like Wikipedia redirect pages and link text. We also envisage using the lexicon to extract Arabic-French parallel corpus from our comparable corpus. The parallel

[3] http://www.larousse.fr/dictionnaires/francais-arabe/

corpus will be used as training data for Statistical Machine Translation.

References

Adafre, S. F. and 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, pages 62–69.

Alexandre Patry and Philippe Langlais. 2011. PARADOCS : Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction inWikipedia. Proceedings of the 4th Workshop on Building and Using Comparable Corpora, 9th Annual Meeting of the Association for Computational Linguistics, Portland, 2011.

Bouma, G., Fahmi, I., Mur, J., G. Van Noord, Van Der, L., and Tiedemann, J. 2006. Using Syntactic Knowledge for QA. In Working Notes for the Cross Language Evaluation Forum Workshop.

Bowker Lynne and Pearson Jennifer. 2002. Working with Specialized Language: A Practical Guide to Using Corpora. Routledge, London/New York.

Brown Peter, F., Pietra, V. J., Pietra, S. A., and Mercer, R. L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. IBM T.J. Watson Research Center, pages 264-311.

Buscaldi D.and Rosso P.. 2006. Mining Knowledge from Wikipedia for the Question Answering Task. Proceedings of the 5th International Conference on Language Resources and Evaluation.

Declerck, T., Perez, A. G., Vela, O., Z., and Manzano-Macho, D. 2006. Multilingual Lexical Semantic Resources for Ontology Translation. In Proceedings of International Conference on Language Ressources and Evaluation (LREC), pages 1492 – 1495.

Erdmann, M., Nakayama, K., Hara, T. Et Nishio, S. 2008. A bilingual dictionary extracted from the wikipedia link structure. In Proceedings of International Conference on Database Systems for Advanced Applications (DASFAA) Demonstration Track, pages 380-392.

Grefenstette, G. 1998. The Problem of Cross-language Information Retrieval. Crosslanguage Information Retrieval. Kluwer Academic Publishers.

Inguna Skadiņa A , Ahmet Aker B , Voula Giouli C , Dan Tufis D , Gaizauskas B , Madara Mieriņa A , Nikos Mastropavlos C. 2010. A Collection of Comparable Corpora for Under-resourced Languages. Fourth International Conference HUMAN LANGUAGE TECHNOLOGIES . “Athena”, Greece.

Isaac Gonzalez Lopez and Pablo Gamallo Otero. 2010. Wikipedia as multilingual source of comparable corpora. In Proceedings of the LREC 2010, Malta.

Ivanova Angelina , 2012. Evaluation of a Bilingual Dictionary Extracted from Wikipedia. InProceedings, 5th Workshop on Building and Using Comparable Corpora (BUCC), Istanbul, Turkey.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In Proceedings of the Human Language Technologies/North American Association for Computational Linguistics, pages 403–411.

Leuf B. and Cunningham W. 2001. The Wiki Way: Collaboration and Sharing on the Internet. Addison-Wesley.

Magdalena Plamada and Martin Volk. 2012. Towards a Wikipedia-extracted Alpine Corpus. 5th Workshop on Building and Using Comparable Corpora at LREC 2012. Istanbul.

Mehdi Mohammadi and Nacer QasemAghaee, 2010. Building Bilingual parallel Corpora based on Wikipedia. In Proceedings of the 2010 Second International Conference on Computer Engineering and Applications - Volume 02, ser. ICCEA ’10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 264–268.

Mihalcea R. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT 2007).

Min-Hsiang Li, Vitaly Klyuev and Shih-Hung Wu. 2010. Multilingual sentence alignment from Wikipedia as multilingual comparable corpora . Proceedings of the 13th International Conference on Humans and Computers. Japan.

Mohammadi M. and QasemAghaee N.. 2010. Building bilingual parallel corpora based on Wikipedia. International Conference on Computer Engineering and Applications, 2:264–268.

Nakayama K., Hara T.and Nishio S. 2007. Wikipedia Mining for An Association Web Thesaurus Construction. Proceedings of the 8th International Conference on Web Information Systems Engineering.

Och, F.J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, pages 19–51, March.

Pablo Gamallo Otero.and Isaac Gonzalez Lopez. 2011. Measuring comparability of multilingual corpora extracted from Wikipedia , Proceedings of Workshop on Iberian Cross-Language NLP tasks (ICL-2011) , September 2011, Huelva, Spain.

Sadat, F. et Terrassa, A. 2010. Exploitation de Wikipédia pour l'Enrichissement et la Construction des Ressources Linguistiques. TALN 2010, Montréal.

Skadiņa, I., Aker, A., Glaros, N., Su, F., Tufiş, D., Verlic, M., Vasiļjevs, A. and Babych, B. 2012. Collecting and Using Comparable Corpora for Statistical Machine Translation, in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey.

Vogel, S., Ney H. and C. Tillmann .1996. HMM-based word alignment in statistical translation. In Preceding of the Conference on Computational Linguistics, pages 836–841, Morristown, NJ, USA.