Measuring Comparability of Multilingual Corpora Extracted from Wikipedia *

Midiendo la comparabilidad de copus multilingües extraídos de la Wikipedia

Pablo Gamallo Otero
Centro de Investigación en Tecnoloxías da Información (CITIUS), Universidade de Santiago de Compostela
Galiza, Spain
pablo.gamallo@usc.es

Issac González López
Cilenis S.L.
Language Engineering Solutions
Santiago de Compostela
Galiza, Spain
isaacjgonzalez@cilenis.com

Resumen: Los corpus comparables son muy útiles en variadas tareas del procesamiento del lenguaje tales como la extracción de léxicos bilingües. Con la mejora de la calidad de los corpus comparables, podemos mejorar la calidad de la extracción. Este artículo describe algunas estrategias para construir corpus comparables a partir de la Wikipedia, y propone una medida de comparabilidad. Fueron realizados algunos experimentos utilizando la Wikipedia portuguesa, española e inglesa.
Palabras clave: Extracción de Información, Corpus Comparables, Léxicos Bilingües, Comparabilidad

Abstract: Comparable corpora can be used for many linguistic tasks such as bilingual lexicon extraction. By improving the quality of comparable corpora, we improve the quality of the extraction. This article describes some strategies to build comparable corpora from Wikipedia and proposes a measure of comparability. Experiments were performed on Portuguese, Spanish, and English Wikipedia.
Keywords: Information Extraction, Comparable Corpora, Bilingual Lexicons, Comparability

1. Introduction

Wikipedia is a free, multilingual, and collaborative encyclopedia containing entries (called "articles") for almost 300 languages (281 in July 2011). English is the more representative one with about 3 million articles. However, Wikipedia is not a parallel corpus as their articles are not translations from one language into another. Many works have been published in the last years focused on its use and exploitation for multilingual tasks in natural language processing: extraction of bilingual dictionaries (Yu y Tsujii, 2009; Tyers y Pieanaar, 2008), alignment and machine translation (Adafre y de Rijke, 2006; Tomás, Bataller, y Casacuberta, 2001), multilingual information retrieval (Pottast, Stein, y Anderka, 2008). There also exists theoretical work analysing symmetries and asymmetries among the different multilingual versions of an entry/article in Wikipedia (Filatova, 2009).

In addition, multilingual articles of Wikipedia have been used as a source to build comparable corpora (Gamallo y González, 2010). The EAGLES - Expert Advisory Group on Language Engineering Standards Guidelines (see `http://www.ilc.pi.cnr.it/EAGLES96/browse.html`) defines a "comparable corpus" as one which selects *similar* texts in more than one language or variety. One of the main advantages of comparable corpora is their versatility to be used in many linguistic tasks (Maia, 2003), like bilingual lexicon extraction (Gamallo y Pichel, 2008; Saralegui, Vicente, y Gurrutxaga, 2008), information retrieval, and knowledge engineering. Besides, they can also be used as training corpus to improve statistic machi-

* This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project OntoPedia, ref: FF12010-14986.

ne learning systems, in particular when parallel corpora are scarce for a given pair of languages. Another advantage concerns their availability. In contrast with parallel corpora, which require (not always available) translated texts, comparable corpora are easily retrieved from the web. Among the different web sources of comparable corpora, Wikipedia is likely the largest repository of similar texts in many languages. We only require the appropriate computational tools to make them comparable.

By taking into account multilingual potentialities of Wikipedia, our main objective is to define a method to measure the similarity (or degree of comparability) of different comparable corpora built from Wikipedia. For this purpose, first we describe some strategies to extract monolingual corpora in Portuguese, Spanish, and English from Wikipedia, by making use of some categories ("Archaeology", "Biology", "Physics", etc.) to make them comparable according to a specific topic. These strategies were described in detail in (Gamallo y González, 2010). Then, we propose a measure of comparability to verify whether the corpora are lowly or highly comparable. For many extraction tasks, such as bilingual lexicon extraction, using highly comparable corpora often leads to better results. There are some works proposing comparability measures between monolingual corpora (Li y Gaussier, 2010; Saralegui y Alegria, 2007), based on the use of existing bilingual dictionaries. However, instead of exploiting dictionaries to compute the comparability degree, we take advantage of the translation equivalents inserted in Wikipedia by means of *interlanguage links*.

This paper is organized as follows. Section 2 introduces two strategies to build comparable corpora from Wikipedia. Next, in Section 3, we propose some comparability measures. Then, Section 4 describe some experiments performed in order to measure the comparability between different corpora built using the strategies defined in Sec. 2 . The last section discusses future tasks that will be implemented in order to extend and improve our tools.

2. Two strategies to Build Wikipedia-Based Comparable Corpora

The input of our strategies is CorpusPedia[1], a friendly and easy-to-use XML structure, generated from Wikipedia dump files. In CorpusPedia, all the internal links found in the text are put in a vocabulary list identified with the tag *links*. In the same way, all the categories (or topics) used to classify each article are inserted in the tag *category*. In addition, there is a tag called *translations* which codifies a list of interlanguage links (i.e., links to the same articles in other languages) found in each article. Categories and translations are very useful features to build comparable corpora. Given these features, we developed two strategies aimed to extract corpora with different degrees of comparability.

Not-Aligned Corpus This strategy extracts those articles in two languages having in common the same topic, where the topic is represented by a category and its translation (for instance, the English-Spanish pair "Archaeology-Arqueología"). It results in a not-aligned comparable corpus, consisting of texts in two languages. We called it "not-aligned" because the version of an article in one language may have not its corresponding version in the other language.

Aligned Corpus The goal is to extract pairs of bilingual articles related by interlanguage links if, at least, one of both contains a required category. It results in a comparable corpus that is aligned article by article.

In Section 4, we will measure the degree of comparability of corpora built by means of these two strategies. Before that, we will define how to measure comparability between Wikipedia-based corpora.

3. Comparability Measures

For a comparable corpus $\mathcal{C}$ of Wikipedia articles, constituted for instance by a Portuguese part $\mathcal{C}_p$ and a Spanish part $\mathcal{C}_s$, a comparability coefficient can be defined on the basis

[1]The software to build CorpusPedia, as well as CorpusPedia files for English, French, Spanish, Portuguese, and Galician, are freely available at `http://gramatica.usc.es/pln/`

of finding, for each Portuguese term t_p in the vocabulary $\mathcal{C}_p^v$ of $\mathcal{C}_p$, its interlanguage link (or translation) in the vocabulary $\mathcal{C}_s^v$ of $\mathcal{C}_s$. The vocabulary of a Wikipedia corpus is the set of "internal links" found in that corpus. So, the two corpus parts, $\mathcal{C}_p$ and $\mathcal{C}_s$, tend to have a high degree of comparability if we find many internal links in $\mathcal{C}_p^v$ that can be translated (by means of interlanguage links) into many internal links in $\mathcal{C}_s^v$. Let $Trans_{bin}(t_p, \mathcal{C}_s^v)$ be a binary function which returns 1 if the translation of the Portuguese term t_p is found in the Spanish vocabulary $\mathcal{C}_s^v$. The binary Dice coefficient, $Dice_{bin}$, between two parts of a comparable corpus $\mathcal{C}$ is then defined as:

$$Dice_{bin}(\mathcal{C}_p, \mathcal{C}_s) = \frac{2 \sum_{t_p \in C_p^v} Trans_{bin}(t_p, \mathcal{C}_s^v)}{|\mathcal{C}_p^v| + |\mathcal{C}_s^v|}$$

We consider that it is not necessary to define the counterpart of the translation function, since the number of ambiguous terms is very low in Wikipedia, and most cases of ambiguity are solved with the so-called "disambiguated pages".

To avoid a bias towards common internal links, that is, towards those links occurring in most articles, we define a specific version of tf_idf weight for each term. In particular, $tf\_idf(t_p)$ is the frequency of term t_p in the Portuguese part of the comparable corpus, multiplied by its inverse *article* frequency in the whole Portuguese Wikipedia. By taking into account the tf_idf of terms, we can define a weighted measure of comparability. Let $Trans_{tf\_idf}(t_p, \mathcal{C}_s^v)$ be a function which returns the smallest value (min) of two tf_idf scores, both $tf\_idf(t_p)$ and $tf\_idf(t_s)$, where t_s is the Spanish translation of t_p in the Spanish part $\mathcal{C}_s$. The weighted Dice coefficient, $Dice_{tf\_idf}$, between two parts of a comparable corpus $\mathcal{C}$ is then defined as follows:

$$Dice_{tf\_idf}(\mathcal{C}_p, \mathcal{C}_s) = \frac{2 \sum_{t_p \in C_p^v} Trans_{tf\_idf}(t_p, \mathcal{C}_s^v)}{\sum_{t_p \in C_p^v} tf\_idf(t_p) + \sum_{t_s \in C_s^v} tf\_idf(t_s)}$$

The experiments described in the next section will be performed with the two comparability measures defined here.

4. Experiments and Results

Taking CorpusPedia as input source, we performed several experiments to build different comparable corpora for three language pairs, namely Portuguese-Spanish, Portuguese-English, and Spanish-English. These corpora were built using the two strategies described in Section 2 and five domain specific seed terms (in the three languages) considered as representative of five domain topics: "Archaeology", "Linguistics", "Physics", "Biology", and "Sport".

Table 1 shows the (binary and tf_idf) Dice scores obtained from measuring the comparability degree of 30 different comparable corpora. For each corpus, the table also shows the size (in Mb) of its two parts. In particular, the first column introduces the two languages of the corpus (pt = Portuguese, sp = Spanish, en = English) and the type of strategy (aligned or not aligned) used to build it. In the second and third columns, we show the two Dice scores. The forth column shows the size of the two parts of the corpus, and the last column contains the two seed terms employed to generate the corpus. In Table 2, we show the Dice scores as well as the size of nine pairs of monolingual corpora randomly generated from Wikipedia.

We can observe first that there are significant differences in terms of comparability between the Dice scores in Table 1 and those obtained from the randomly generated monolingual pairs in Table 2. It follows that corpora built by means of our strategies (not aligned and aligned) are actually *comparable*. Then, we should note that in the comparable corpora of Table 1, the Dice scores based on tf_idf are about 70 % higher than those based on the binary function. By contrast, in randomly generated corpora (Table 2), there are no significant differences between $Dice_{bin}$ and $Dice_{td\_idf}$. It means that our tf_idf makes the Dice similarity score higher if the two evaluated corpus parts are actually comparable.

As it was expected, not-aligned corpora tend to be larger than the aligned ones. However, if we just compare the smallest parts of each corpus, the differences are not very important: the smallest parts of not-aligned corpora are only 15 % larger than those of aligned corpora. This is in accordance with the fact that aligned corpora are more balanced in terms of size, since no part is much larger than the other one. As far the corpus size is concerned, let us note that, in average, English parts are clearly larger than the Spanish ones, which are slightly larger than the Portuguese ones. In general, English ar-

Corpora	Dice (bin)	Dice (tf-idf)	Size (in Mb)	Seed terms
pt-sp (not aligned)	.068	.086	0.6Mb/3.4Mb	Arqueologia, Arqueología
pt-en (not aligned)	.041	.067	0.6Mb/8.4Mb	Arqueologia, Archaeology
sp-en (not aligned)	.090	.140	0.4Mb/8.4Mb	Arqueología, Archaeology
pt-sp (aligned)	.179	.199	0.4Mb/0.2Mb	Arqueologia, Arqueología
pt-en (aligned)	.127	.140	0.4Mb/1.1Mb	Arqueologia, Archaeology
sp-en (aligned)	.181	.226	2.0Mb/2.9Mb	Arqueología, Archaeology
pt-sp (not aligned)	.078	.129	0.8Mb/1.7Mb	Linguística, Lingüística
pt-en (not aligned)	.054	.136	0.8Mb/5.1Mb	Linguística, Linguistics
sp-en (not aligned)	.074	.170	1.7Mb/5.1Mb	Lingüística, Linguistics
pt-sp (aligned)	.140	.214	0.6Mb/0.8Mb	Linguística, Lingüística
pt-en (aligned)	.128	.194	0.5Mb/1.2Mb	Linguística, Linguistics
sp-en (aligned)	.150	.257	0.9Mb/1.7Mb	Lingüística, Linguistics
pt-sp (not aligned)	.200	.374	4.4Mb/4.8Mb	Física, Física
pt-en (not aligned)	.123	.287	4.4Mb/12Mb	Física, Physics
sp-en (not aligned)	.270	.403	4.8Mb/12Mb	Física, Physics
pt-sp (aligned)	.237	.390	3.6Mb/4.7Mb	Física, Física
pt-en (aligned)	.178	.348	3.8Mb/11Mb	Física, Physics
sp-en (aligned)	.220	.387	3.4Mb/7.6Mb	Física, Physics
pt-sp (not aligned)	.130	.227	2.4Mb/1.5Mb	Biologia, Biología
pt-en (not aligned)	.102	.193	2.4Mb/9.4Mb	Biologia, Biology
sp-en (not aligned)	.068	.129	1.5Mb/9.4Mb	Biología, Biology
pt-sp (aligned)	.197	.328	1.6Mb/2.8Mb	Biologia, Biología
pt-en (aligned)	.186	.308	1.8Mb/4.5Mb	Biologia, Biology
sp-en (aligned)	.213	.294	0.9Mb/1.3Mb	Biología, Biology
pt-sp (not aligned)	.083	.148	11Mb/35Mb	Desporto, Deporte
pt-en (not aligned)	.026	.085	11Mb/333Mb	Desporto, Sport
sp-en (not aligned)	.047	.136	35Mb/333Mb	Deporte, Sport
pt-sp (aligned)	.175	.266	9.7Mb/15Mb	Desporto, Deporte
pt-en (aligned)	.189	.334	11Mb/20Mb	Desporto, Sport
sp-en (aligned)	.206	.290	20Mb/29Mb	Deporte, Sport
pt-sp (not aligned)	.111	.192	3.8Mb/9.3Mb	**Overall**
pt-en (not aligned)	.069	.153	3.8Mb/73Mb	**Overall**
sp-en (not aligned)	.109	.195	9.3Mb/73Mb	**Overall**
pt-sp (aligned)	.185	.279	3.2Mb/4.7Mb	**Overall**
pt-en (aligned)	.161	.264	3.5Mb/7.6Mb	**Overall**
sp-en (aligned)	.194	.290	6.2Mb/8.5Mb	**Overall**

Cuadro 1: Dice similarity between several comparable corpora in Portuguese, Spanish, and English.

Corpora	Dice (bin)	Dice (tf-idf)	Size (in Mb)
pt-sp1 (random)	.012	.012	2.2Mb/0.9Mb
pt-en1 (random)	.003	.003	2.2Mb/0.4Mb
sp-en1 (random)	.003	.003	0.9Mb/0.4Mb
pt-sp2 (random)	.016	.014	1.5Mb/3.0Mb
pt-en2 (random)	.017	.014	1.5Mb/42Mb
sp-en2 (random)	.017	.015	3.0Mb/42Mb
pt-sp3 (random)	.008	.006	0.2Mb/0.5Mb
pt-en3 (random)	.001	.001	0.2Mb/1.4Mb
sp-en3 (random)	.005	.005	0.5Mb/1.4Mb

Cuadro 2: Dice similarity between randomly generated pairs of monolingual corpora.

ticles tend to have more words than Spanish and Portuguese articles. As it was suggested by one of the reviewers of the article, one of the reasons for the difference in size in the case of aligned corpora is that Spanish and Portuguese entries seem to be summaries of the English ones. So, to increase comparability between an aligned pair of articles, the longer article could be shortened by removing those parts which are not present in the other language, obtaining, this way, a more comparable pair of articles.

Finally, as it was expected, aligned corpora are significantly more comparable (i.e., higher Dice coefficient) than not-aligned corpora. In average, $Dice_{td\_idf}$ increases 80 % the comparability of aligned-corpora with regard to not-aligned ones. So, considering that aligned corpora only decreases 15 % in size in relation to not-aligned corpora, we can conclude that the aligned strategy seems to be more appropriate to build comparable corpora from Wikipedia.

5. Conclusions and Future Work

The emergence of multilingual resources, such a Wikipedia, makes it possible to design new methods and strategies to compile corpus from the web, methods that are more efficient and powerful than the traditional ones. In particular, the semi-structured information underlying Wikipedia turns out to be very useful to build comparable corpora. In this article, we proposed two strategies to build comparable corpora from Wikipedia and a way to measure their degree of comparability. The experiments led us to conclude that corpora aligned article by article are more comparable than not aligned corpora. Besides, they consist of two balanced corpus parts in terms of size. Finally, they are not much smaller than not aligned corpora.

In future work, we will be focused on how to improve the strategies to build comparable corpora by extending coverage (more articles) without losing comparability. For this purpose, we will test and evaluate techniques to expand categories using a list of similar terms identified as hyponyms or co-hyponyms of the source category. In order to find hyponyms and co-hyponyms of a term, it will be required to build an ontology of categories using the semi-structured information of Wikipedia (Chernov et al., 2006; Ponzetto y Navigli, 2009; de Melo y Weikum, 2010). On the other hand, we will evaluate comparability in an indirect way. In particular, we will use the generated corpora on tasks requiring comparable corpora as input (e.g., bilingual lexicon extraction). The better the extracted lexicon, the more comparable the input corpus should be. Finally, we believe that our method for aligning pairs of articles could be useful for related tasks, such as Wikipedia infoboxes alignment in different languagues (Adar, Skinner, y Weld, 2009).

Bibliografía

Adafre, S.F. y M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. En *11th Conference of the European Chapter of the Association for Computational Linguistics*, páginas 62–69.

Adar, Eytan, Michael Skinner, y Daniel S. Weld. 2009. Information arbitrage across multi-lingual wikipedia. En *Second ACM International Conference on Web Search and Data Mining , WSDM*.

Chernov, Sergey, Tereza Iofciu, Wolfgang Nejdl, y Xuan Zhou. 2006. Extracting semantic relationships between wikipedia categories. En *SemWiki2006 - From Wiki to Semantics*, Budva, Montenegro.

de Melo, Gerard y Gerhard Weikum. 2010. Menta: inducing multilingual taxonomies from wikipedia. En *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, páginas 1099–1108.

Filatova, Elena. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. En *CLEAWS3*, páginas 30–37, Colorado.

Gamallo, Pablo y Isaac González. 2010. Wikipedia as a multilingual source of comparable corpora. En *LREC 2010 Workshop on Building and Using Comparable Corpora*, páginas 19–26, Valeta, Malta.

Gamallo, Pablo y José Ramom Pichel. 2008. Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary. *LNCS*, 4919:413–423.

Li, Bo y Eric Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. En

20th International Conference on Computational Linguistics (COLING 2010, páginas 644–652.

Maia, Belinda. 2003. What Are Comparable Corpora. En *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, páginas 27–34, Lancaster, UK.

Ponzetto, Simone Paolo y Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating wikipedia. En *Proceedings of the 21st international jont conference on Artifical intelligence*, páginas 2083–2088.

Pottast, M., B. Stein, y M. Anderka. 2008. A wikipedia-based multilingual retrieval model. En *Advances in Information Retrieval*, páginas 522–530.

Saralegui, X. y I. Alegria. 2007. Similitud entre documentos multilíngües de carácter científico-técnico en un entorno Web. En *Procesamiento del Lenguaje Natural*, página 39.

Saralegui, X., I. San Vicente, y A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. En *LREC 2008 Workshop on Building and Using Comparable Corpora*.

Tomás, J., J. Bataller, y F. Casacuberta. 2001. Mining Wikipedia as a Parallel and Comparable Corpus. En *Language Forum*, volumen 1, página 34.

Tyers, M.F. y J.A. Pieanaar. 2008. Extracting Bilingual Word Pairs from Wikipedia. En *LREC 2008, SALTMIL Workshop*, Marrakesh, Marocco.

Yu, Kun y Junichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. En *Machine Translation Summit XII*, Ottawa, Canada.