Turning Wikipedia into a resource for language research

Alberto Montero-Asenjo, Carlos A. Iglesias

Grupo de Sistemas Inteligentes (GSI)
Universidad Politécnica de Madrid, Spain
{amontero,cif}@gsi.dit.upm.es

Abstract

Wikipedia is a valuable resource whose usage goes beyond the encyclopedia itself. In this paper the proposal is to use Wikipedia as a large source of text, suitable for language research, explaining the followed procedure to turn Spanish Wikipedia raw data into a suitable text source, considering the format of source data (wiki syntax), the conversion from written text to individual sentences or the conversion from acronyms or numbers to the way they are said. The case explained here is specific in some parts to the Spanish wikipedia, but the ideas and some steps of the followed procedure can be generalised to any language or text source.

1. Introduction

Language resources (corpus) are usually collected and distributed by dedicated organisations and the cost to the public is usually high or it has restrictions on their applicability. Examples of such corpus for Spanish are CREA [1], *Corpus de la lengua española contemporánea* [2], Argentina [3] or ARTHUS [4]. Up to some extent, price and availability restrictions are caused by the fact that usually this kind of resources are built every time from the ground, and efforts to collect, revise and tag (at several levels) the corpus is huge. But there already are large sources of text data, some of those even free (as in free speech, not as in beer) that could be potentially used to build larger and better databases for language research, without big investments.

Wikipedia [5] is collaborative effort to build a free multilingual encyclopedia. Its name is a portmanteau of the words wiki (a type of collaborative website) and encyclopedia. It was launched in 2001 by Jimmy Wales and Larry Sanger and currently is operated by the non-profit Wikimedia Foundation. It is one of the largest, fastest growing and most popular general reference work currently available on the Internet (accorgind to Wikipedia webpage [5]).

As of December 2007, Wikipedia had approximately 9.25 million articles in 253 languages, comprising a combined total of over 1.74 billion words for all Wikipedias. The English Wikipedia edition passed the 2,000,000 article mark on September 9th 2007, and as of 21 January 2008 it had over 2,185,000 articles consisting of over 950,000,000 words. Wikipedia's articles have been written collaboratively by volunteers around the world, and the vast majority of its articles can be edited by anyone with access to the Internet. Having steadily risen in popularity since its inception, it currently ranks among the top ten most-visited websites worldwide (these figures have been taken from [6]).

Spanish Wikipedia [7] is a much smaller project than the English one. It was founded a few months later than the general project (on May 2001) and at the beginning of 2008 it had more than 300,000 articles and more than 600,000 user from most of the Spanish-speaking countries.

Wikipedia has been used in other scenarios than the encyclopedic search, and it has been previously used as a research resource in fields like semantic research, knowledge extraction or natural language processing [8, 9, 10, 11, 12], where knowledge embedded in Wikipedia was the most valuable resource. This paper proposes not using the knowledge but the text expressing that knowledge, as a representation of language, and up to some extent, speech.

Wikipedia main strengths are its nature of free resource, and thus available to anyone, and its big size (as stated before, Spanish Wikipedia has more than 300,000 articles), thus allowing for a wide variety of words, topics and writing styles. The main weakness is the unsupervised nature, thus not ensuring quality, and requiring some quality control and improvements steps.

Next section (2) is fully devoted to explain the procedure to turn Wikipedia raw data into useful text, section 3 refers to public availability of the generated resources, section 4 draws the main conclusions derived from this work and section 5 is about future lines.

2. Data processing

Data processing consists in a set of steps to convert Wikipedia data into useful text. Figure 1 represents an overview of the process. Basically it has four steps: convert wiki markup to plain text, split paragraphs into sentences (being aware of certain aspects), rewrite sentences as they would be read and, finally, remove incorrect words from vocabulary (and sentences having those words).

2.1. Data source

Every few months, Wikipedia is dumped to a large XML file per language and made publicly available. A dump from Spanish Wikipedia dated by 06/07/2007 was used as raw source data. This dump is not currently available as old dumps are removed (every dump requires about 1Gb of disk space). Latest dumps can be found at [13]. For the processing of the raw XML dump, Perl module Parse::MediaWikiDump (available at CPAN [14]), and data was splitted into articles storing each one in a separate file.

A sample piece of text taken from article "Tebas (Grecia)"("Tebas (Greece)") will be used to illustrate the followed

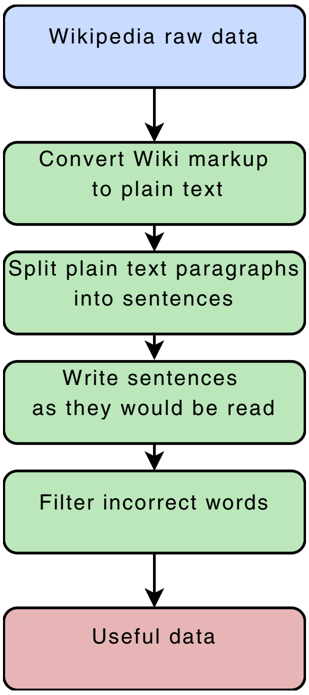
Wikipedia raw data
Convert Wiki markup to plain text
Split plain text paragraphs into sentences
Write sentences as they would be read
Filter incorrect words
Useful data

Figure 1: Data processing overview

process, showing the transformations suffered by the text on every step.

Sample
En la actualidad, el lugar de la antigua ciudadela, [[Cadmea]], se encuentra ocupado por la ciudad de Thíva ("&Theta;&#942;&beta;&alpha;") que fue reconstruida después del [[terremoto]] de [[1893]]. La ciudad actual tiene 24.400 habitantes ([[2001]]), llamados "tebanos"..

2.2. From wiki markup to plain text

Wikipedia XML dumps have data as users wrote it, so it has not only the useful text but many other symbols and references corresponding to wiki syntax. This extra markup must be removed in order to extract valuable text. For this task, Perl module Text::MediawikiFormat (available at CPAN [14]). Despite the simpleness of the wiki markup, there are inevitable syntactic errors not parseable by Text::MediawikiFormat, so it was necessary an extra filtering stage to remove, for example, unmatched brackets or extra '='. This step was also used to beauty text by removing extra spaces and other minor changes. The effect on the sample text denoted above is shown below.

Sample
En la actualidad, el lugar de la antigua ciudadela, Cadmea, se encuentra ocupado por la ciudad de Thíva ($\Theta\eta\beta\alpha$) que fue reconstruida después del terremoto de 1893. La ciudad actual tiene 24.400 habitantes (2001), llamados tebanos.

2.3. Sentences division

Since the corpus is intended to be applied to speech recognition, it is needed to convert it into sentences, instead of written formated text, such as paragraphs, lists, text in parenthesis, enumerations after semicolons and others. These cases have to be addressed in order to use that text. The following points may serve as an example of the followed approach:

- Abbreviations or acronyms are translated into the full words, in order to be process the corpus as a speech corpus.
- Paragraphs were divided into sentences. Dots are the main sentence separator (as well as line or paragraph end), but with some considerations like dots being number separators or part of an acronym (processed previously).
- Text inside parenthesis is consider as different sentences, thus generating two. The first one is the original one without the text inside parenthesis and the second one the text inside parenthesis.

Sample
En la actualidad, el lugar de la antigua ciudadela, Cadmea, se encuentra ocupado por la ciudad de Thíva que fue reconstruida después del terremoto de 1893. $\Theta\eta\beta\alpha$. La ciudad actual tiene 24.400 habitantes , llamados tebanos. 2001.

2.4. Sentences as they would be read

Written text and spoken speech are closely related, but they are not the same. As an example one may consider numbers (either in arabic or roman format). In written text "2001" may appear, while in spoken speech it will be said as "two thousand and one". But there are many other examples, as mathematical operations, where "+" must be replaced by "plus" or acronyms, which are usually spoken by spelling letters. Finally, capital letters were converted to lower case and commas and other unrecognised symbols were removed. After this step the previous example will remain as follows:

Sample
en la actualidad el lugar de la antigua ciudadela cadmea se encuentra ocupado por la ciudad de thíva que fue reconstruida después del terremoto de mil ochocientos noventa y tres la ciudad actual tiene veinticuatro mil cuatrocientos habitantes, llamados tebanos dos mil uno

2.5. Vocabulary filtering

After the steps explained above a huge amount of text was available. Main figures are shown below.

Number of articles: 69,541

Sentences: 3,280,428

Vocabulary size: 549,962

The dynamic range of the histogram of occurrences is so high that a singe picture cannot show all information. The word having most occurrences is ’de’ (in English *of*, *from*), appearing 2,526,038 times. Near half a million words appear less than 30 times. Figure 2 shows a crop of the histogram of vocabulary, considering only words appearing less than 50 times, covering almost 94% of the words. Horizontal axis represents number of occurrences and the vertical one the number of words having that number of occurrences.

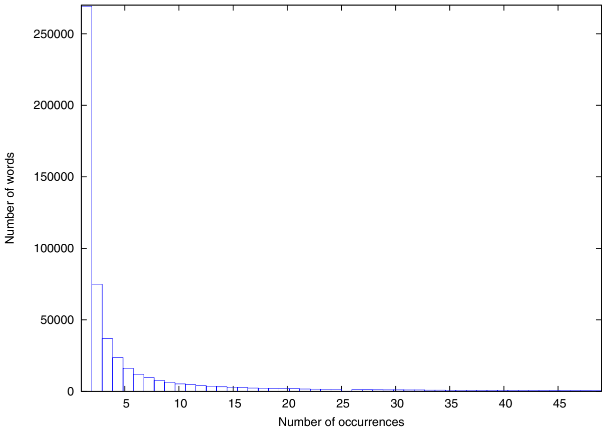

Figure 2: Word occurrences histogram before vocabulary cleaning

The dictionary of Real Academia Española (RAE, the official Spanish language authority), in the 2001 edition [15] has about 90.000 entries, considering near a 10% of archaisms, not including neither all verbal forms nor plural and gender dependent forms for nouns and adjectives. Despite the fact that many words can be composed by adding prefixes and suffixes to standard words, given this figures, more than half a million words seems a huge number.

Due to misspelled words and foreign terms (as propper names, technical words or etymological terms, to cite only a few) it was expected to have a large amount of words, where correct words will occur many times and incorrect ones only a few, making a simple threshold-based decision good enough. But reality is that correct and incorrect words are much more coupled in number of occurrences than expected. This situations makes vocabulary filtering and cleaning a difficult task.

The filtering process consisted in an iterative process of data examination and word removal. When a word was removed, all the sentences where it appeared were removed. Articles with no sentences were erased. The followed heuristics have been defined:

1. Remove words with only one occurrence. These words are considered to be foreign words, misspelled ones or too rare in common Spanish.
2. Remove words with double consonants (bb, cc, dd, ...) and less than 3 occurrences.
3. Remove words with only one occurrence.

Step		Articles	Sentences	Words
Intial	Abs	69,541	3,280,428	549,962
1	Abs	69,541	2,998,212	275,196
	%	100	91.4	50.04
2	Abs	68,679	2,981,025	263,943
	%	98.76	90.87	47.99
3	Abs	68,679	2,959,262	241,180
	%	98.76	90.21	43.85
4	Abs	68,650	2,958,498	241,105
	%	98.72	90.19	43.84
5	Abs	51,925	1,294,040	114,068
	%	74.67	39.45	20.74

Table 1: Remaining data evolution after filtering stages

4. Remove words with the same letter repeated 3 o more times consecutively (aaa, bbb, ccc, ...)

This steps, despite the simplicity, greatly reduced the amount of selected data, keeping only a half of the original vocabulary (241,105 remaining words).

The reduction of available data was very large, but a closer look to the remaining words revealed that there were many terms not valid in Spanish, but difficult to discriminate based on occurrences. A more powerful filtering scheme was needed, and it was achieved by manual inspection of words, identifying words and word patterns not present in Spanish and removing them. For example there is no Spanish words ending with ’-ly’ (typical in English adverbs) and words ending with ’-lae’ or ’-mae’ or starting with ’phy-’ are usually Latin words found in technical terms (as species names).

At this point other problems were revealed, not related to words but to character encoding. Wikipedia is primarily UTF-8 encoded, but we found many words having other codification schemes (ISO 8859 1) which made the filtering process a bit harder and we took the decision to remove non UTF-8 encoded words (altough some of them may have been correctly re-encoded automatically and preserved).

Manual revision of near a quarter of million words is a very expensive and time consuming process, so we decide to achieve it iteratively, inspecting a subset on each iteration. The benefits of this approach was that after removing a word and its associated sentences, other potential candidates for removal are automatically removed, thus decreasing the total number of words to inspect. After 8 iterations, over 20,000 words and word patterns were identified and removed from vocabulary.

5. Iteratively inspect vocabulary and select words and word patterns to remove.

Table 1 and figure 3 summarizes the effect of the different filtering stages. In the table you can see the absolute amounts of reamining data and the percentages of the initial dat, while the plot shows the evolution of percentages. As you can see, the most aggresive stage is the last one.

The most common word is still ’de’ appearing 367,013 times, and 31,729 words appear only once in the whole remaining text. Figure 4 shows again a crop of the histogram of the remaining data restricted to words appearing less than 50 times, which represents 93% of the words.

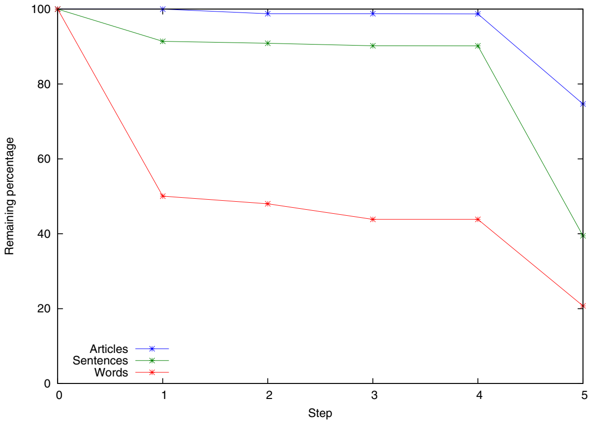

Figure 3: Remaining data evolution after filtering stages

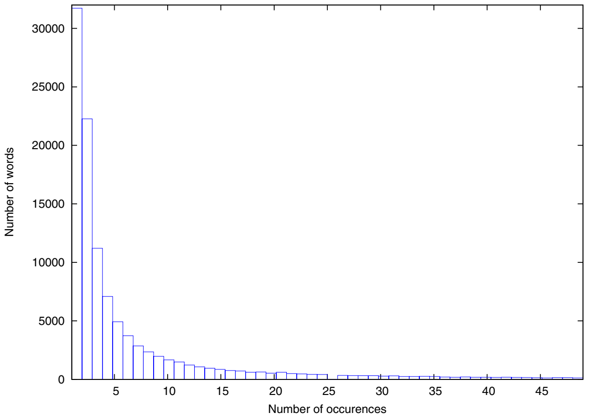

Figure 4: Word occurrences histogram after vocabulary cleaning

3. Data availability

As mentioned above, this work starts from free resources and has made use of many free software tools. To maintain this spirit and to allow others to make improvements and new researches, the generated data as well as the involved scripts have been made public and can be found at `http://wp4lr.sourceforge.net/`. All material can be used and redistributed under the same terms as Wikipedia itself. Any suggestion, improvement or bug detection will be welcomed.

4. Conclusions

Wikipedia is a large source of data but in a format that is not the most usual in the language research community. This paper presents an effort to make that source a valuable resource.

The has been no measures about the goodness of the generated data, as it is difficult to make meaningful comparisons. What other source can be compared to Wikipedia in terms of size and topics covered? How to make such a comparison? Perplexity measures perhaps? And, what should be measured at the end, the data or the process? We have prefer to keep the original goal that was to have a large amount of text for our current researches.

The process to utilise Wikipedia can be automated and reutilised across languages and text sources up to some extent, but the most difficult and time consumming step (vocabulary filtering) is Spanish specific.

Due to the fact of public availability and continuous improvement of articles, it could be expected to have a mid to high quality resource. But the fact is that many misspelled and badly encoded words were found. Some criticisms have been published about the quality of Wikipedia at semantic level (accuracy, political and ideological bias), and the results found here may be a source for another level of criticism. This fact along with the unsupervised nature of the proposed procedure makes necessary a quality control.

For the purpose of this article, initial quality of data may have reduced the amount of required work and increased the size of available data (as sentences with unappropriated words were fully removed), but even after the extensive filtering stage the amount of data is still very large and suitable for the purposes it was conceived.

5. Future lines

One of the weak points of the followed procedure is related to the vocabulary filtering stage, as it is quite expensive, language specific and its quality is difficult to assess. A way to improve these aspects may be the use of already made dictionaries, to move out sentences containing words not covered by the external vocabulary. Obviously this dictionary has to be large enough as to cover all (or at least mostly) of the vocabulary present in the Wikipedia. Such kind of resources are available for Spanish [1, 16], but as the number of queries that have to be made (more than half a million) is quite large, the process must carefully designed to not overload those sites, and count with the agreement of the site.

6. References

[1] R. A. E. B. de datos (CREA) [on line], “Corpus de referencia del español actual,” http://corpus.rae.es/creanet.html, 2008.

[2] S. LABORATORIO DE LINGÜÍSTICA INFORMÁTICA, Universidad Autónoma de Madrid, “Corpus de referencia de la lengua española contemporánea,” http://www.lllf.uam.es/corpus/corpus.html.

[3] “Corpus lingüístico de referencia de la lengua española en argentina,” http://www.lllf.uam.es/~fmarcos/informes/corpus/coarginl.html.

[4] S. Universidad de Santiago, “Archivo de textos hispánicos de la universidad de santiago,” http://gramatica.usc.es/EspArthus.html.

[5] J. Wales and L. Sanger, “Wikipedia, the free encyclopedia,” http://www.wikipedia.org, 2001.

[6] “Wikipedia entry for Ẅikipedia,;” http://en.wikipedia.org/wiki/Wikipedia.

[7] “Spanish wikipedia,” http://es.wikipedia.org.

[8] T. Zesch, I. Gurevych, and M. Mühlhäuser, “Analyzing and Accessing Wikipedia as a Lexical Semantic Resource,” in *Data Structures for Linguistic Resources and*

Applications, G. Rehm, A. Witt, and L. Lemnitzer, Eds. Tuebingen, Germany: Gunter Narr, Tübingen, 2007, pp. 197–205.

[9] T. Zesch and I. Gurevych, "Analysis of the Wikipedia Category Graph for NLP Applications," in *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, 2007, pp. 1–8.

[10] T. Zesch, I. Gurevych, and M. Mühlhäuser, "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)*, 2007, pp. 205–208.

[11] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. New York, NY, USA: ACM, 2007, pp. 41–50.

[12] S. P. Ponzetto and M. Strube, "Deriving a large-scale taxonomy from wikipedia," in *AAAI*. AAAI Press, 2007, pp. 1440–1445.

[13] "Spanish Wikipedia dumps," http://download.wikimedia.org/eswiki/latest/.

[14] "Comprehensive Perl Archive Network," http://www.cpan.org.

[15] "Spanish RAE dictionary figures," http://buscon.rae.es/draeI/html/drae/cifras.htm.

[16] "Corpus del español," http://www.corpusdelespanol.org/.