La Wikipedia como fuente multilingüe de corpus comparables

En este artículo se describe un método automático de selección de corpus comparables a partir de la Wikipedia, utilizando categorías temáticas como elementos restrictivos. Nuestra estrategia se fundamenta en dos propiedades de la Wikipedia: el ser un recurso multilingüe y el tratarse de una enciclopedia libre disponible para descarga en formato XML. Las herramientas y los corpus generados dispondrán de licencia libre GPL (General Public License).
Palabras clave: Corpus comparables, extracción de información multilingüe

This article describes an automatic method to select comparable corpora from Wikipedia using categories as topic restrictions. Our strategy is based on two properties of Wikipedia: to be a multilingual resource and to be a free encyclopedia available in a XML file. Tools and corpus will be distributed under GPL license (General Public License).
Key words: Comparable corpora, multilingual information extraction

1. Introducción

La Wikipedia es una enciclopedia libre multilingüe *online* y colaborativa con entradas para alrededor de 300 lenguas, de las cuales, el inglés es la más representativa con casi 3 millones de artículos. Como se puede observar en la tabla 1, el número de entradas/artículos en las lenguas más usadas de la Wikipedia alcanza ya un nivel más que suficiente para poder llevar a cabo investigación multilingüe con solvencia. La tabla muestra que el español se encuentra en noveno lugar, con 460 mil entradas, muy cerca del portugués, que alcanza las 470 mil. Nuestras primeras experiencias se realizarán con estas dos lenguas.

En consonancia con la enorme expansión y rápido crecimiento de la Wikipedia, en los últimos años han surgido numerosos trabajos que explotan este recurso para diferentes objetivos multilingües: extracción de diccionarios bilingües (Yu & Tsujii 2009; Tyers & Pienaar 2008), alineamiento, paralelización y traducción automática (Adafre & Rijke 2006; Tomás et al. 2008), recuperación de información multilingüe (Potthast et al. 2008). Por último, comienzan a aparecer trabajos sobre la comparabilidad de los artículos en diferentes lenguas de la Wikipedia, con la posibilidad de elaborar, a partir de ellos, corpus comparables (Filatova 2009).

Lenguas	Número de artículos
Inglés	2.826.000
Alemán	888.000
Francés	786.000
Polaco	593.000
Italiano	556.000
Japonés	576.000
Holandés	528,000
Portugués	470.000
Español	460.000
Ruso	376.000

Tabla 1: *Presencia de las lenguas en la Wikipedia. Las 10 primeras lenguas ordenadas en función del número de artículos (datos de abril 2009).*

Un corpus comparable se compone de textos en diferentes lenguas con una temática similar (McEnery & Xiao 2007). Una selección de artículos de periódicos en diferentes lenguas, tratando del mismo tema y en la misma época supone un buen ejemplo de corpus comparable. Este tipo de corpus es útil en múltiples y variadas líneas de investigación. Por citar sólo algunas, puede servir para ayudar a realizar estudios contrastivos, o bien como base para la extracción automática de léxicos y terminologías bilingües, así como fuente de entrenamiento de sistemas de traducción automática (cuando hay falta de corpus paralelos en el par de lenguas objetivo del sistema). Una de las principales características de los corpus comparables es su enorme expansión. A diferencia de los corpus paralelos , que exigen la traducción de una lengua a otra, los corpus comparables se encuentran, de forma natural, en la web, a medida que aumenta el número de lenguas presentes en este medio.

De hecho, como ya ha sido mencionado antes, la Wikipedia es una fuente natural de corpus comparables. Sólo es preciso construir las herramientas adecuadas para extraerlos.

Aprovechando las potencialidades multilingües de la Wikipedia, el objetivo de este artículo es describir un método para extraer de ella corpus comparables en función de dos parámetros de variabilidad: las lenguas concretas que se quieran escoger y el tema seleccionado. En concreto, dadas dos lenguas y un tema, nuestra estrategia crea un corpus con textos en las dos lenguas escogidas que versan sobre la temática seleccionada. Las herramientas y los corpus generados con ellas dispondrán de licencia libre GPL (*General Public License*) y estarán disponibles para descarga en: *http://gramatica.usc.es/pln*

Este artículo se organiza del siguiente modo. La sección 2 describe el modo en que transformamos la Wikipedia en un corpus codificado, que llamamos CorpuPedia. La sección 3 presenta las estrategias de construcción de corpus comparables a partir de la CorpusPedia. En la sección 4, se ofrecen datos empíricos de la CorpusPedia, así como de algunas experiencias realizadas a partir de las estrategias definidas en 3. El artículo se cierra con algunas consideraciones acerca de las nuevas tareas que pretendemos llevar a cabo para ampliar y mejorar nuestras herramientas.

2. CorpusPedia

La primera parte de nuestro método consiste en transformar los ficheros fuente de la Wikipedia en un conjunto de ficheros con un formato de fácil manipulación: la CorpusPedia. Para ello desarrollamos herramientas que permiten descargar automáticamente la Wikipedia en los idiomas requeridos y aplicar después un proceso de conversión del XML descargado al XML formato del corpus.

2.1. *Formato de la Wikipedia*

La Wikipedia es descargable en su totalidad en ficheros XML, en distintas versiones en la que se puede escoger que cantidad de metadatos descargar. Como se puede ver en el ejemplo de una entrada de la Wikipedia que se muestra en la figura 1; la forma de acceso a los datos es también simple y muy eficiente. La diferencia respecto a la web al uso es que el texto en vez de tener un formato plano, html o xhtml, tiene un formato propio de la Wikipedia, que ha de ser convertido a texto sin formato.

En la figura 1 podemos ver el extracto de una entrada de la Wikipedia en formato XML, si nos fijamos en el campo text, podremos observar el formato específico del que hablábamos.

```
<page>
    <title>Arqueoloxía</title>
    <id>3</id>
    <revision>
      <id>1310468</id>
      <timestamp>2009-10-06T02:42:14Z</timestamp>
      <contributor>
        <username>SieBot</username>
        <id>2109</id>
      </contributor>
      <minor />
      <comment>bot Engadido: [[ku:Arkeolojî]]</comment>
      <text xml:space="preserve">{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]]   e   [[obxecto]]s   da   [[antigüidade|
antigüidade]], especialmente a través dos seus restos. O nome ven
do  [[lingua  grega|grego]]  ''archaios'',  &amp;quot;vello&amp;quot;  ou
&amp;quot;antigo&amp;quot;,     e     ''logos'',     &amp;quot;ciencia&amp;quot;,
&amp;quot;saber&amp;quot;.

[...]

[[zh:□□□]]
[[zh-yue:□□]]</text>
    </revision>
  </page>
```

Figura 1: *Ejemplo del formato XML de la Wikipedia (artículo gallego "Arqueoloxía").*

2.2. *Formato de la CorpusPedia*

El formato de la CorpusPedia consiste, esencialmente, en un título y el texto de cada entrada, junto con otras informaciones que se extraen gracias al formato semiestructurado de la Wikipedia y de ciertas convenciones entre los editores (Clark et al. 2009).

En la figura 2 podemos observar el código XML del corpus generado a partir de la Wikipedia. Los campos *title*, *category* y *plaintext* son los necesarios relativos al uso de este fichero como corpus comparable, siendo el apartado *category* el que puede aportar información sobre la temática del texto y así agruparlo con otros textos de categoría similar en caso de búsqueda de alta comparabilidad. El campo *wikitext* contiene el formato original de la Wikipedia, que puede ser útil para futuras extracciones y del cual, aplicando un parser, se obtiene el campo *plaintext* (i.e., texto plano sin codificación).

```
<article>
      <title>Arqueoloxía</title>
      <category>Arqueoloxía</category>
      <related>Antropoloxía, Arqueoloxía industrial, Arqueoloxía
submarina</related>
      <links>ciencia, arte|artes, monumento|monumentos, obxecto,
antigüidade|antigüidade, lingua grega|grego, cultura, estudo,
psicolóxico, condutistas, antropoloxía, idade de pedra, Idade
Media, Arqueoloxía industrial, Antropoloxía, Arqueoloxía
industrial, Arqueoloxía submarina</links>
      <translations># Arqueologia Arqueología Archaeology
      Archéologie Arqueologia Arkeologia # Archeologia
      Archeologie Археология Αρχαιολογία</translations>
      <plaintext>A arqueoloxía é a ciencia que estuda as artes,
monumentos e obxectos da antigüidade, [...] o que se coñece como
Arqueoloxía industrial.</plaintext>
      <wikitext>{{Historia en progreso}}

A '''arqueoloxía''' é a [[ciencia]] que estuda as [[arte|artes]],
[[monumento|monumentos]] e [[obxecto]]s da [[antigüidade|
antigüidade]], [...]
[...]
[[yi:ארכעאלאגיע]]
[[zh:□□□]]
[[zh-yue:□□]]</wikitext>
</article>
```

Figura 2: *Formato del corpus generado a partir de la Wikipedia*

El campo *translations* es una lista de los enlaces *interlanguage*, es decir un enlace a esa misma entrada pero en otro idioma; lo que aporta una herramienta muy útil para crear comparabilidad como veremos más adelante. Esta lista de enlaces siempre está ordenada del mismo modo (gl pt es en fr ca eu al it cs bg el). Además, en caso de no existir el enlace, se coloca "#" para indicar explícitamente la no existencia.

Existen otros dos campos que aportan más información y más relaciones con otras entradas de Wikipedia. El campo *related* aporta el título de otras entradas relacionadas con la actual, que han sido así explicitadas en Wikipedia. Por otro lado, *links* es el conjunto de enlaces salientes a otras entradas.

3 Estrategias para crear corpus comparables

Dada la estructura y la información contenida en la CorpusPedia, es factible, no sólo, agrupar artículos sobre una misma temática, sino también relacionarlos con artículos que traten la misma temática en otras lenguas. Por ello, la estructura de la CorpusPedia permite construir con cierta comodidad corpus comparables. Para llevar a cabo esta tarea, creamos varias herramientas orientadas a extraer corpus con diferentes grados de comparabilidad. Estas tres herramientas, que se corresponden con tres estrategias, se describen a continuación.

3.1 *Comparables sin alinear*

Esta estrategia extrae los artículos en dos lenguas que tratan de un mismo tema, donde el tema es sugerido por medio de una categoría y su traducción (por ejemplo el par *Arqueología-Arqueologia*, en castellano y portugués). En concreto, el algoritmo es el siguiente:

```
Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2,
donde C2 es una traducción de C1 en L2, se procede a:
```

- extraer todos los artículos de la corpuspedia de L1, siempre y cuando C1 esté dentro de la lista de categorías de cada artículo procesado;

Repetir el proceso partiendo de los artículos de L2.

El resultado es un corpus comparable no-alineado, compuesto por textos en dos lenguas (L1 y L2) que abordan la misma temática: C1 o C2. Es un corpus no-alineado porque el título de un artículo en una lengua puede o no tener su traducción en la otra lengua, es decir, puede o no tener un enlace *interlanguage* a un artículo de la otra lengua.

3.2. *Alineamiento estricto*

El corpus resultado del anterior proceso puede ser visto como demasiado heterogéneo, pues incluye artículos en una lengua que no tienen su correspondiente versión en la otra. Por ejemplo, puede haber un artículo español titulado "Arqueología de España" que no tiene un enlace *interlanguage* en portugués, es decir un artículo sin su correspondiente versión "Arqueologia de Espanha". El método que utilizamos para construir un corpus alineado a nivel de los artículos, permitiendo sólo recoger aquellos que tienen enlaces *interlanguage* en la otra lengua, es el siguiente:

Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2, donde C2 es una traducción de C1 en L2, se procede a:

- extraer todos los artículos de L1 de la corpuspedia, siempre y cuando: i) C1 esté dentro de la lista de categorías de cada artículo procesado, ii) cada artículo contenga un enlace *interlanguage* a un artículo de L2 conteniendo la categoría C2.

Repetir el proceso partiendo de los artículos de L2 y eliminar las inconsistencias.

El resultado es un corpus comparable alineado de manera muy estricta, ya que no sólo cada

artículo en una lengua tiene su artículo correspondiente en la otra, sino que ambos artículos comparten la misma restricción categorial.

3.3. *Alineamiento laxo*

El alineamiento estricto puede dejar fuera artículos relevantes, por ejemplo, aquellos que, aun teniendo un enlace *interlanguage* a un artículo de la otra lengua, no cumplen la restricción categorial. En concreto, puede haber artículos categorizados en la Wikipedia española como siendo de "Arqueología", pero que no fueron categorizados en portugués por su correspondiente bilingüe "Arqueologia". De hecho, la versión portuguesa de la Wikipedia está menos categorizada que la española. Esta escasez categorial resta cobertura a la estrategia descrita en la subsección anterior (3.2). Para subsanar esta circunstancia, proponemos otro método de alineamiento más laxo. El objetivo es extraer todos los artículos que tienen enlaces *interlangue* conteniendo la categoría requerida en, al menos, una de las dos lenguas. El algoritmo es el siguiente:

> Dadas dos lenguas, L1 y L2, y dos categorías, C1 y C2, donde C2 es una traducción de C1 en L2, se procede a:
>
> - extraer todos los artículos de L1 de la corpuspedia, siempre y cuando: i) C1 esté dentro de la lista de categorías de cada artículo procesado, ii) cada artículo contenga un enlace *interlanguage* a un artículo de L2.
>
> - extraer todos los artículos de L2 que tienen enlace con los artículos extraídos en el proceso anterior.
>
> Repetir los dos procesos desde L2 y eliminar los artículos duplicados y las inconsistencias.

El resultado es un corpus alineado artículo a artículo, aunque de temática no tan específica

como la conseguida por la estrategia más restrictiva.

4. Experiencias y Resultados

En el momento actual, el texto plano (*plaintext*) de la CorpusPedia es de aproximadamente 20 millones de palabras para la versión en gallego de la Wikipedia, 120 millones en la versión portuguesa y 180 en la versión en español. El hecho de que la versión española tenga más palabras que la portuguesa contrasta con el mayor número de artículos introducidos en esta última (ver tabla 1 en la introducción). De ello se infiere que los artículos de la versión portuguesa tienden a ser más pequeños que los de la española.

Tomando como base la CorpusPedia, realizamos una experiencia para construir corpus comparables español-portugués sobre una temática específica, arqueología, utilizando las tres estrategias descritas en la sección anterior. La temática específica de los textos españoles fue seleccionada por medio de la categoría "Arqueología", y se usó su traducción, "Arqueologia", para seleccionar los textos portugueses. La tabla 2 muestra los resultados numéricos de la experiencia.

Estrategia	**Tamaño (en palabras)**	**Número de artículos**
es/pt no-alineado	344.000 / 64.000	420 / 100
es/pt alineado estricto	27.000 / 11.000	19 / 19
es/pt alineado laxo	132.000 / 64.000	119 / 119

Tabla 2: Tres estrategias para construir corpus comparables español-portugués usando la categoría "Arqueología-Arqueologia".

La tabla 2 deja entrever que hay bastante disparidad en el tamaño de los corpus. Por

ejemplo, usando la estrategia básica sin alineamiento, el corpus español alcanza las 344 mil palabras frente a las 64 mil del portugués. Esto se debe, sobre todo, a que se han encontrado 420 artículos en español restringidos por la categoría "Arqueología", frente a los 100 en portugués asociados a la categoría "Arqueologia". Además, el tamaño de los artículos es también significativamente mayor en el corpus español. Usando la estrategia de alineamiento laxo, para los mismos artículos (119), el corpus español alcanza las 130 mil palabras, frente a las 64 mil del portugués. De aquí se deduce que los artículos en portugués sobre arqueología tienden a contener la mitad de información que los españoles.

Finalmente, la tabla 3 muestra, a modo de ejemplo, 5 pares bilingües de títulos correspondientes a la lista de artículos extraídos usando la estrategia "alineado estricto". Esto nos permite observar el grado de comparabilidad de los corpus extraídos, si bien una cuantificación del grado de comparabilidad será objeto de estudio de posteriores experimentos.

Artículos en español	**Artículos en portugués**
Arqueoastronomía	Arqueoastronomia
Arqueología	Arqueologia
Arqueología bíblica	Arqueologia bíblica
Arqueología procesual	Arqueologia processual
Arqueología subacuática	Arqueologia subaquática

Tabla 3: 5 primeros artículos extraídos usando la estrategia de alineado estricto.

5. Conclusiones y trabajo futuro

La aparición de recursos multilingües, como la Wikipedia, posibilitan nuevos métodos de creación de corpus a partir de la web, que son más eficientes y potentes que los tradicionales. Asimismo permiten crear corpus comparables a partir de la extracción de la(s) temática(s) de cada porción de texto. En la actualidad, estamos buscando cómo mejorar las estrategias de extracción ampliando la cobertura de los artículos seleccionados sin perder precisión. Para ello, estamos evaluando dos técnicas: por un lado, el uso del campo *related* para aumentar el número de categorías restrictivas, por otro, la expansión automática por hipónimos de la categoría escogida. Esta última estrategia sólo se podrá llevar a cabo si se dispone de una ontología de categorías previamente construida. Una de nuestras tareas, en la actualidad, es construir una ontología de categorías a partir de información estructurada de la Wikipedia.

Para la ampliación de cobertura hemos formulado algunas estrategias: el uso de los enlaces internos de wikipedia (no sólo los enlaces *interlanguage*) que ya son extraídos y el uso de enlaces externos que generarán corpus con alguna de las técnicas de *web as corpus*. Si bien estas estrategias ampliarían la cobertura, probablemente disminuirían la comparabilidad, por lo que esos valores deberán ser evaluados y posteriormente incorporado o no el nuevo corpus, dependiendo de las necesidades de comparabilidad y cobertura.

Por último, realizaremos evaluaciones sobre el grado de comparabilidad de los corpus generados. Para ello, utilizaremos métodos inspirados en los trabajos de Saralegi & Alegria (2007).

Bibliografía

Adafre, S.F. & de Rijke, M. (2006) "Finding Similar Sentences acorss Multiple Languages in Wikipedia", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 62-69.

Clark M, Ian Ruthven & Patrik O'Brian Holt (2009), "The Evolution of Genre in Wikipedia", *JLCL,* vol 24 (1), (1-22).

Filatova, Elena (2009) "Directions for Exploiting Asymmetries in Multilingual Wikipedia", *Proceedings of CLEAWS3*, Boulder, Colorado, pp. 30-37.

González, I. and Gamallo, P. (2010) "Estrategias para la elaboración de corpus comparables a partir de la web ", XXXIX Simposio internacional de la SEL

McEnery, A. M. and Xiao, R. Z. (2007) "Paralell and comparable corpora: What are they up to?" In: James, G. and Anderman, G., (eds.) *Incorporating Corpora: Translation and the Linguist.* Translating Europe . Multilingual Matters, Clevedon, UK.

Potthast, M. Stein, B. and Anderka, M. (2008) "A Wikipeda-Based Multilingual Retrieval Model".

Saralegi X. and Alegria I. (2007), "Similitud entre documentos multilíngües de carácter científico-técnico en un entorno Web", *Procesamiento del Lenguaje Natural,* 39.

Tomás, J., Bataller, J. and Casacuberta, F. (2008) "Mining Wikipedia as a Parallel and Comparable Corpus", *Language Forum*, 34(1).

Tyers, M.F. and Pieanaar, J.A. (2008) "Extracting bilingual word pairs from Wikipedia", LRE 2008, SALTMIL Workshop, Marrakech, Marroco.

Yu, Kun and Tsujii, Junichi (2009). Bilingual Dictionary Extraction from Wikipedia. *Proceedings of MT Summit XII*, Ottawa, Canada.