GaIn: un buscador Internet/Intranet avanzado para textos en euskera

Ibon Aizpurua, Iñaki Alegria, Nerea Ezeiza
Ixa Taldea. Informatika Fakultatea (Euskal Herriko Unibertsitatea)
acpalloi@si.ehu.es

Resumen

En este artículo se presentan las tareas realizadas para combinar la explotación de la información de la Web y las técnicas del NLP dando como resultado un buscador avanzado para textos en euskera, todo ello desarrollado dentro del proyecto *GaIn* financiado parcialmente por la Diputación Foral de Gipuzkoa y por el programa Universidad-Empresa del Gobierno Vasco (UE-1999-2). Este trabajo ha sido desarrollado por el grupo *IxA* (ixa.si.ehu.es) de la Universidad del País Vasco en colaboración con el portal Jalgi (www.jalgi.com) de Plazagune S.L.

La herramienta realizada es un buscador de Internet/Intranet con su robot, indexador y buscador, que tiene dos módulos de NLP que lo convierten en avanzado: por un lado un identificador de idioma y por otro un lematizador robusto.

I. La herramienta

Los componentes básicos de un buscador (*search engine*) son tres: robot, indexador y buscador. En lugar de programarlo hemos recurrido al software disponible (sobre todo al software libre). Una vez estudiadas las herramientas libres de estas características nos decidimos por Swish-E de la universidad de Berkeley, ya que se adaptaba perfectamente a las características requeridas: modularidad, completitud, libre distribución y disponibilidad de código fuente multiplataforma y multiformato. Se ha hecho la adaptación en un servidor Web Linux-Apache.

Robot

Es el módulo que va buscando textos por la red y seleccionando los que pueden ser interesantes. Solamente hemos tenido que hacer unos pequeños retoques con el objetivo de: integrar más formatos, distinguir varios tipos de documento entre ellos los *frames*, ofrecer la posibilidad de salir del dominio, saltar los *proxies*

Además le hemos añadido un identificador de idioma que actúa como filtro para solo indexar textos en euskera. Como nuestra primera necesidad era solamente la respuesta binaria si el texto estaba en euskera o no, hemos seguido las técnicas de las palabras más frecuentes y los trigramas y el algoritmo ha sido muy simple y basado en umbrales de frecuencia en función de unos primeros resultados experimentales. El sistema se ha evaluado probándolo con textos obtenidos de Internet en función del idioma y los resultados son muy buenos. En ningún caso un texto en otro idioma se identificó como en euskera y solamente en un caso de 30 se ha tomado por otro idioma el de un texto en euskera (página con nombres de grupos de música).

Indexador

El indexador es el módulo encargado de generar unos índices para que la posterior búsqueda sea lo más eficiente posible. Para que los índices no sean demasiado grandes se eliminan del índice las palabras funcionales de aparición frecuente. Una de las labores realizadas ha sido generar una lista de este tipo para el euskera.

Sin embargo, el cambio más importante ha sido la indexación por lemas con vistas a cumplir el objetivo enunciado. Hemos integrado un lematizador (Ezeiza et al., 98) del que disponíamos previamente y conseguido, además de un buscador más avanzado, un índice más compacto. Para conseguir que el lematizador sea robusto, el proceso de análisis se hace incrementalmente en tres fases: 1) lematización estándar, 2) lematización de variantes dialectales y errores frecuentes y 3) lematización sin léxico. Más del 99% de las palabras se lematizan correctamente. En el futuro queremos aumentar la inteligencia del buscador indexando solo los lemas o sintagmas más significativos, siguiendo el trabajo que en curso sobre extracción de terminología (Alegria *et al.*, 99).

Buscador

Es el modulo que por medio de una interfaz captura las palabras clave en que el usuario basa su pregunta y consultado los índices genera una página con enlaces a las páginas que contienen la información correspondiente.

En el módulo de búsqueda se ha adaptado la interface y se ha incluido también el módulo de lematización en previsión de que se pregunte por la forma y no por el lema.

II. Demostración

El resultado obtenido es totalmente satisfactorio para un dominio Intranet. Para su utilización como buscador genérico de Internet debe ser mejorado para una mayor eficiencia, y además se debe poder tener un ancho de banda mayor del que disponemos. Sin embargo, si el robot lo apoyamos sobre una primera prebúsqueda en uno o varios buscadores comerciales los resultados son buenos aunque asumiremos las limitaciones de los buscadores base.

Esta herramienta se esta utilizando en dominios Intranet, destacando dos casos:

- el portal Jalgi <www.jalgi.com> que incorpora este buscador tanto a contenidos propios (Megadenda, enciclopedia, etc.) como a otros indexados a partir de información recuperada por el robot.
- La hemeroteca del diario Egunkaria <http://www.egunkaria.com/egun1999/sarrera.html> que indexa los contenidos completos, desde enero de 1999, del único diario en euskera.

En la Figura 1 vemos un ejemplo de la búsqueda en este segundo caso. Se ha buscado *uholdeen ondorioa* (el efecto de las inundaciones) y en las dos primeras opciones de la respuesta en el titular del artículo aparecen ...*uholde baten ondorioz*... (como consecuencia de una riada) y ...*uholdeen ondorioz* (como consecuencia de las inundaciones).

Bibliografía

Alegria I., Ezeiza N., Oronoz M., Urizar R. (1999) *Extracción Automática de Terminología a partir de Etiquetado y Lematización*. VI Simposio Internac. de Comunicación Social. Cuba, 1999.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. (1998) *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98, Montreal (Canada). August 10-14, 1998.

Grefenstette, G. (1995) *Comparing Two Language Identification Schemes*. Proc. of 3rd International. Conf. on Statistical Analysis of Textual Data. Italy. 1995

Swish-E http://sunsite.berkeley.edu/SWISH-E

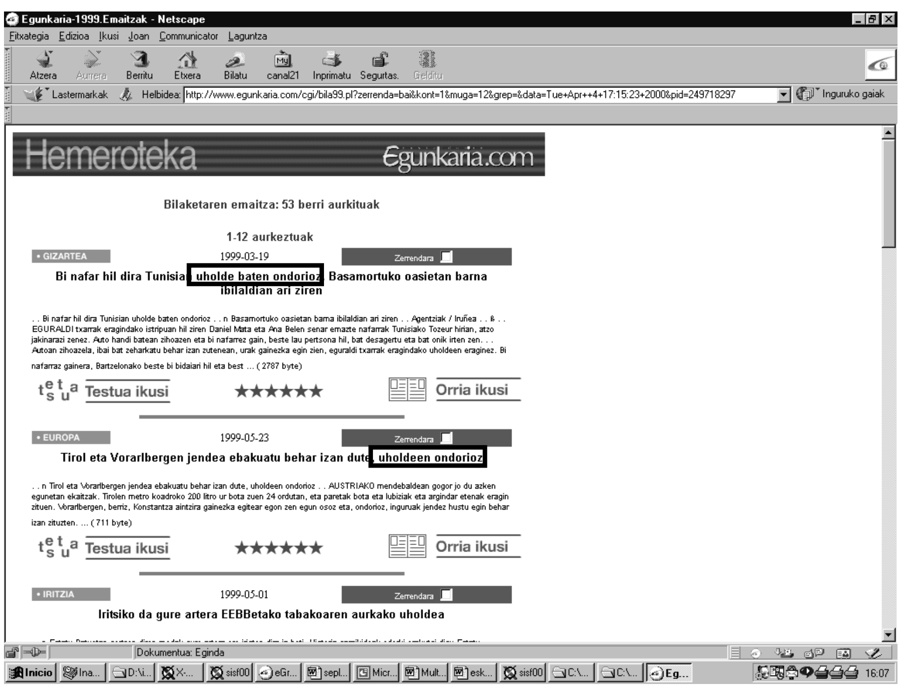

Fig.1: Resultado de la búsqueda de *uholdeen ondorioa*