IÑAKI ALEGRIA (i.alegria@si.ehu.es)
NEREA EZEIZA
MAITE ORONOZ
RUBEN URIZAR

Informatika Fakultatea, Universidad del País Vasco.
P.K. 649. 20080 Donostia. Euskal Herria.

Extracción Automática de Terminología a partir de Etiquetado y Lematización

Abstract

In the last few years, several tools for automatic terminology extraction from corpora have been developed by different research groups. Although the process is almost automatic, human intervention is still required in order to make the final selection from the terms be automatically chosen. As an example we can cite some tools: LEXTER, Termight from AT&T, TERMS from IBM and NPtool.

The areas of application of these tools can be divided into two main groups: terminological work and information indexing. A combination of NLP techniques (based on linguistic knowledge) and statistical techniques is usually used to select the terms from a corpus, though in some cases the first ones are not applied. Nevertheless, in complex inflected languages like Basque, poor results will be obtained if only the surface aspect of words is dealt with. More complex linguistic treatment as semantics is proposed for normalisation of terminology.

Linguistic techniques are basically used to make the initial selection of terms. Morphosyntactic models are usually used in the process, so it is advisable to have the text already analysed or at least labelled. A shallow syntactic process should be convenient as well.

In our project, we use a morphology based lemmatiser/tagger which uses a combination of statistical and linguistic information to assign a tag and a lemma to each word, so that we will be able to obtain and apply statistical values based on word-forms or lemmas. Nowadays, some tasks are in progress in the construction of the extractor of terminology for Basque: modelling of term structures, building a shallow syntactic analyser and programming the statistical module.

A survey based on bibliography about terminology extraction and the description of the components of our tool are the main sections in this paper.

1 Introducción

En los últimos años se vienen desarrollando herramientas para la extracción automática de terminología. Como resultado de los avances en las investigaciones en el área del Procesamiento del Lenguaje Natural (NLP) y otras áreas afines, la tecnología actual esta en disposición de desarrollar este tipo de herramientas aunque, vistos los resultados obtenidos, todavía es necesario un postproceso manual para proceder a una última selección de los términos representativos de un texto a partir del resultado del extractor automático.

Los campos de aplicación de estas herramientas son variados pero se pueden dividir en dos grandes grupos: trabajo terminológico propiamente dicho, como puede ser la construcción de listas o diccionarios de términos, e indexación de información, que esta estrechamente ligada con todo el área de recuperación de la información (*Information Retrieval*). Respecto al primer campo cabe decir que la obtención (semi)automática de terminología es de gran interés en los campos de traducción técnica y en el mundo editorial. Además, en áreas en donde el desarrollo terminológico es muy dinámico, como puede ser el mundo de la informática, sin herramientas de este tipo es casi imposible un trabajo terminológico efectivo. El campo de la indexación de la información presenta un interés incluso superior, debido, en gran parte, al fenómeno Internet y a la gran cantidad de información disponible en línea pero que no va acompañada con herramientas de búsqueda/selección adecuadas.

Algunos autores (El-Hadi & Jouis, 98) distinguen entre terminología aplicada, que referencia a la que tiene una aplicación concreta (indexación, traducción, etc.), con la genérica que sería de un uso general. Sólo en el último caso parece adecuado que el resultado sea una lista de términos que componen el texto, ya que en el primer caso la herramienta debe estar adaptada a la aplicación particular.

Como ya hemos dicho los resultados son prometedores para el inglés o francés, pero todavía queda trabajo por realizar y más para otras lenguas con menos recursos a su disposición. Al mismo tiempo se está discutiendo sobre la forma de evaluar y comparar los resultados de las diferentes herramientas.

En nuestro caso queremos desarrollar una herramienta de este tipo para el euskera (lengua vasca), lengua de flexión rica y aglutinante, lo que conlleva una necesidad mayor de análisis morfosintáctico. Además al ser el proceso de unificación incipiente (las primeras normas datan de 1968), aunque se ha desarrollado un gran esfuerzo, no falto de polémica, en fijar del léxico básico, las decisiones e investigaciones a nivel de terminología son escasas.

El proyecto que presentamos está en fase de desarrollo y, aunque no tenemos resultados concretos, en el artículo se hace, además de una amplia revisión bibliográfica, una descripción de las partes ya en funcionamiento (etiquetador, lematizador) así como de la arquitectura del sistema completo y de las tareas en fase de realización.

2 Extracción automática de terminología

Haciendo un repaso de la bibliografía se detecta en los últimos años un creciente interés por este tema. Como ejemplo de ello podemos nombrar una serie de herramientas: *LEXTER* (Bourigault, 92), *Termight* de AT&T (Church & Dagan, 94), *TERMS* de IBM (Justeson & Katz, 95), *NPtool* (Arpper, 95) y ACABIT (Daille &Jacquemin, 98).

La primera tarea a afrontar es la delimitación de las características de los términos. La mayoría de los trabajos se suelen limitar a buscar sintagma nominales o algunas de las formas morfosintácticas más frecuentes de los mismos. El trabajo de Justeson & Katz estudia para el inglés estas estructuras y llega a la conclusión anterior, aunque no esta claro que esto sea cierto para otros idiomas. De todas formas siempre hay que tener en cuenta que muchos términos son multipalabra con lo que el etiquetado morfosintáctico será una técnica previa para poder localizarlos.

Una vez estudiada la estructura de los términos objetivo de la búsqueda se suele proceder a construir la herramienta para extracción automática de los hipotéticos términos, ya que posteriormente habrá, en la mayoría de los casos, un postproceso manual para la selección definitiva.

Como veremos a continuación, para construir las herramienta de extracción de terminología se suelen combinar técnicas de NLP (basadas en conocimiento lingüístico) con técnicas estadísticas más o menos sofisticadas.

Aunque muchas de estas herramientas se quedan en la mera obtención de una lista cada vez son más numerosas las propuestas que intentan procesar esta lista y el corpus original y obtener resultados más elaborados como pueden ser clasificar la lista de términos, obtener redes terminológicas, desambiguar términos, etc. En estos casos el tratamiento lingüístico se suele complicar y el tratamiento semántico adquiere una especial relevancia.

2.1 Técnicas lingüísticas

Las técnicas lingüísticas se utilizan preferencialmente para una preselección de los posibles términos a tratar. Estos precandidatos suelen ser posteriormente sometidos a algún filtro de carácter cuantitativo para obtener la lista definitiva de candidatos. Para la primera preselección conviene tener el texto lo más analizado posible, siendo normalmente un requerimiento mínimo el etiquetado morfosintáctico (Cutting *et al.*, 92) (Brill, 92). Estas herramientas conocidas como *taggers* han tenido un gran desarrollo durante los últimos años y están disponibles para un número importante de idiomas.

Unido al proceso de etiquetado morfosintáctico, está la lematización, que aunque en las lenguas de nuestro entorno (español, francés, ingles, ...) a veces no se le da gran importancia, en lenguas de flexión nominal rica, y más en las aglutinantes, es de gran importancia para conseguir unos buenos resultados. De todas formas los resultados no son sencillos de interpretar ya que p. ej. en euskera *sistema eragileak* o sistema eragilearen son flexiones en nominativo plural y genitivo del término *sistema eragile* (sistema operativo), sin embargo *sistemaren eragile* (accionador del sistema) no corresponde al mismo término.

Aunque debido a razones de eficiencia un análisis sintáctico profundo no suele aplicarse, un análisis sintáctico superficial si es necesario cuando los términos se quieren seleccionar en función de patrones sintácticos.

De todas formas en ciertos proyectos se evitan tanto el tratamiento morfológico como sintáctico y se recurre sencillamente a conjuntos de dos y tres palabras (Su *et al.*, 96). En algunos sistemas también se utiliza una lista de delimitadores de términos (artículos, preposiciones, etc.) para evitar el tratamiento sintáctico.

Además de para la preselección, el tratamiento lingüístico es importante también tanto para la normalización lingüística de los términos como para un tratamiento posterior, intentándose detectar, para ello, casos de flexión (esto se puede tener en cuenta en la preselección), variación gráfica o sintáctica del mismo término, sinonimia, hiperonimia, etc., aunque para muchos de estos casos ya es imprescindible disponer de una herramienta de tratamiento semántico. Las últimas referencias bibliográficas reportan trabajos en este sentido: (Daille &Jacquemin, 98) (Hamon & Nazarenko, 98) (Maynard & Ananiadou, 98).

2.2 Técnicas estadísticas

En la mayoría de los proyectos se utilizan técnicas estadísticas para restringir entre los precandidatos que ya cumplen los requisitos lingüísticos aquellos que deben considerarse términos del texto.

Los métodos aplicados varían muchos según los proyectos: lo más simple es la exigencia de una frecuencia mínima (Justeson & Katz, 95), pero lo más habitual es la combinación de diferentes fórmulas estadísticas entre la que destaca por su utilización la llamada *mutual information* (Church & Hanks, 90). Por esta fórmula se establece el índice de correlación entre dos componentes de un hipotético término de la siguiente forma:

$$MI(a,b) = \log_2 \left(P(a,b) / P(a)\,P(b) \right)$$

donde P(a) y P(b) son las probabilidades de aparición de cada uno de los componentes en el corpus y P(a,b) la probabilidad de que aparezcan conjuntamente.

A partir de los resultados de esta fórmula se pueden hacer otros cálculos como sucede en (Su *et al.*, 96) que compara los resultados sobre el corpus de estudio con resultados generales obtenidos a partir de un gran corpus equilibrado, para obtener las probabilidades relativas que pueden ser más significativas que las absolutas a la hora de discriminar la terminología.

También se pueden aplicar técnicas estadísticas para la normalización de términos como sucede en el proyecto de la Universidad de Manchester (Frantzi & Ananiadou, 96) donde se propone el uso de una fórmula de nombre C-value, que es capaz de distinguir que *soft contact lenses, hard contact lenses,* y *contact lenses* son términos relacionados y que sin embargo *soft contact* no.

Como ya se ha apuntado el uso de herramientas semánticas deben mejorar los resultados que se obtienen con estas fórmulas. Así Ferrari y Prince (Ferrari & Prince, 96) advierten de la limitación del uso de la fórmula de *mutual information* si no se tiene en cuenta fenómenos como la sinonimia. De todas formas según (Maynard & Ananiadou, 98) usando características semánticas de WordNet no se mejora significativamente el reconocimiento de términos, aunque se apunta que puede ser debido a razones técnicas y no de principio.

2.3 Resultados

Los resultados que se obtienen no son suficientes para una extracción totalmente automática pero si permiten ver la viabilidad de la realización de buenas herramientas de ayuda en la extracción de terminología. A la hora de evaluar las herramientas se utilizan las medidas de cobertura o *recall* (nº de términos detectados / nº total de términos) y precisión (nº de términos correctos / nº total de resultados). Como se ve en le resumen de la tabla 1 al ajustar uno de los parámetros se suele desajustar el otro, con lo que en la mayoría de los proyectos se opta por obtener una buena cobertura a pesar de perder precisión (esto se suele corregir con un postproceso manual). Los resultados parecen converger hacia una cobertura de aproximadamente el 95% con una precisión cercana al 50%.

De todas formas esta forma de evaluación de las herramientas, consistente en comparar los resultados con una lista previa extraída manualmente, es puesta en cuestión por varios autores

(Bourigault & Habert, 98) (Daille & Jacquemin, 98), remarcándose que para una evaluación conveniente es importante tener en cuenta la aplicación para la que está destinada la herramienta.

En la Tabla 1 también se pueden apreciar las características más reseñables de las herramientas estudiadas siendo una muestra del estado del arte en el campo de la extracción de terminología.

CARACTERÍSTICA	LEXTER (Bourigault, 92)	TERMS (Justenson, 95)	(Frantzi & Ananiadou, 96)	(Su *et al.*, 96)
PRESELECCIÓN	sintáctico muy superficial	sintáctico	sintáctico *(adj|nom)*nom*	digramas o trigramas
ETIQUETADO	SI	SI	SI (Brill)	SI
LEMATIZACIÓN	NO	NO	NO	SI
BASE ESTADÍSTICA	frecuencia	frecuencia	frecuencia	mutual information
OTRAS ESTADÍSTICAS	NO	NO	C-value	frecuencia relativa
SEMÁNTICA	NO	NO	NO	NO
COBERTURA/ PRECISIÓN	%95 / ??	?? / %85	%93/%45 %30/%82	%96/%48 %97/%40

Tabla 1.- Características de varios sistemas

3 Aplicación al euskera

En el grupo IXA (http://ixa.si.ehu.es/) de la Universidad del País Vasco estamos desarrollando un extractor de terminología para el euskera (lengua vasca). Durante los últimos años, en proyectos anteriores, hemos desarrollado un analizador/generador morfológico (Alegria *et al.*, 96) que es la base de un lematizador/etiquetador (Aduriz *et al.*, 96). El análisis morfológico se realiza en tres fases incrementales correspondientes al análisis estándar, el análisis de errores comunes y variantes dialectales y el análisis de términos cuyo lema no se encuentra en el léxico. En las tres fases se usa el mismo programa (basado en la morfología de dos niveles) pero con distinto conocimiento lingüístico a nivel de léxico y de reglas morfofonológicas.

También se puede destacar el modulo de desambiguación del lema asociado al análisis de palabras cuyo lema no está en el léxico, ya que por cada etiqueta distintos lemas son posibles y no siempre el más corto es el más conveniente.

El lematizador/etiquetador combina métodos estocásticos con conocimiento lingüístico (Ezeiza *et al.*, 98). Es de resaltar la robustez de estas herramientas que son capaces de analizar y etiquetar y lematizar cualquier palabra aunque su lema no aparezca en el léxico lo que nos permite tratar cualquier texto sin necesidad de actualizar el léxico.

Basándonos en esas herramientas estamos desarrollando varios trabajos en paralelo de cara a la construcción del extractor de terminología:

1) modelización de la composición de los términos. A partir de 3 diccionarios técnicos (administración, informática y fútbol) estamos cuantificando la estructura morfosintáctica de los términos. Disponemos de los primeros resultados y aunque los tenemos que estudiar más a fondo cabe reseñar que los sintagmas nominales no cubren todo el espectro ya que los verbos tienen una cierta frecuencia, y que dentro de los sintagmas nominales deben detectarse casos de declinación concretos y nominalizaciones. Así queda clara la importancia del proceso morfosintáctico previo, como también es resaltado por Daille y Jacquemin (Daille & Jacquemin, 98) cuando resaltan la importancia de la morfología derivacional para obtener buenos resultados.

2) realización de un analizador sintáctico superficial que nos permite buscar los modelos de los términos. Para ello usamos un modelo de estados finitos (para conseguir una eficiencia razonable) que permita la búsqueda basándose en expresiones regulares. Previamente hemos desarrollado un reconocedor de fechas, magnitudes y locuciones comunes.

3) realización del modulo estadístico que maneje tanto las estadísticas a nivel local como a nivel comparativo. Posteriormente, y en función de los resultados de los experimentos, decidiremos si aplicar un aparato estadístico más complejo.

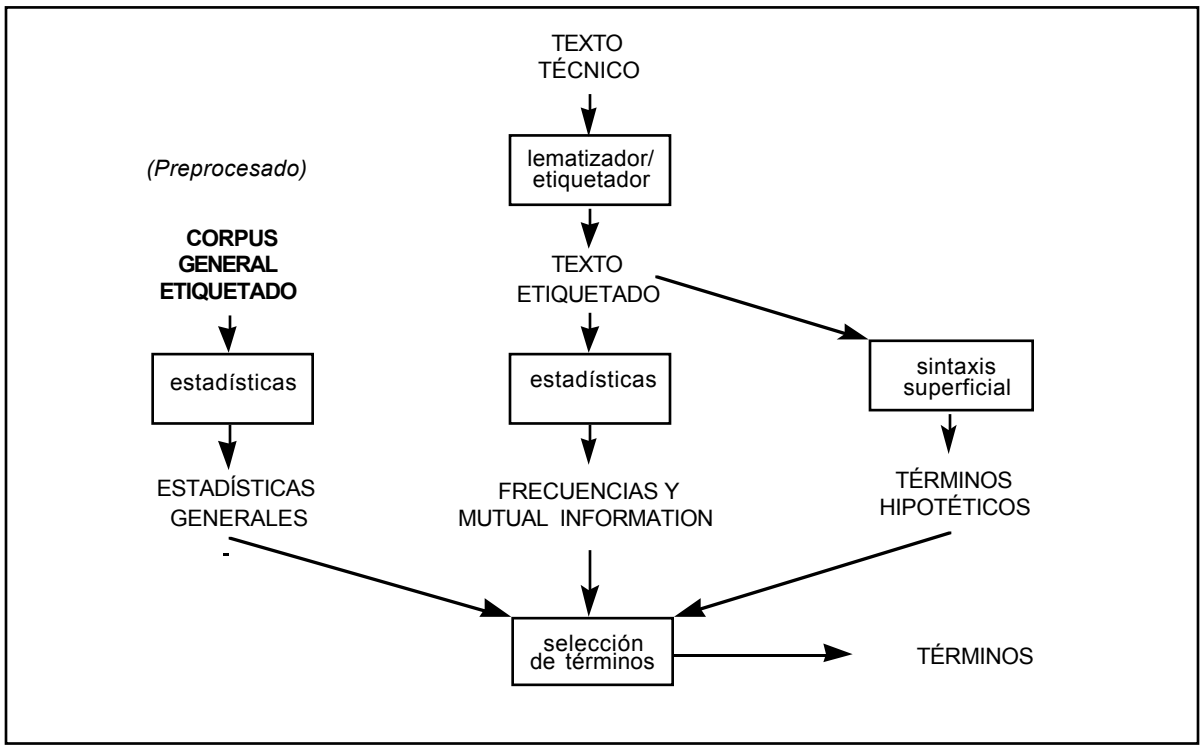

Figura 1.- Arquitectura del extractor de terminología propuesto

La arquitectura del sistema que vamos a construir se refleja en la figura 1, y por el momento prescinde del tratamiento semántico ya que, por ahora, no disponemos de ninguna herramienta para su análisis. Según se puede ver en dicha figura, a partir de un gran corpus equilibrado se van a obtener estadísticas de referencia de palabras, lemas y correlaciones; que nos valdrán para relativizar las frecuencias que se obtengan de los textos a procesar. Así podremos distinguir las palabras y términos multipalabra de frecuencia no habitual que serán los candidatos a formar parte de la terminología.

Teniendo realizado el preproceso anterior, cuando a partir de un texto (libro, artículo, manual, etc.) se quiere obtener la terminología que incluye, se procederá al análisis morfológico y posterior lematización/etiquetado del texto, seguido del análisis sintáctico superficial. A partir de ello se obtienen los precandidatos que seguirán un conjunto de patrones sintácticos determinados y se pasará a su tratamiento estadístico, tanto a nivel interno del texto por medio del calculo de correlaciones, usando la formula de *mutual information*, como a nivel global por la comparación con las estadísticas generales. Basándose en estos se obtendrá la lista de candidatos a ser considerados términos propios del texto analizado.

Reconocimientos:

El presente trabajo está parcialmente subvencionado por el Departamento de Cultura del Gobierno Vasco y por la comisión interministerial de Ciencia y Tecnología (Proyecto de Investigación Item, TIC96-1234-C03-02)

Bibliografía:

(Alegria et al., 96) Alegria I., Artola X., Sarasola K., Urkia M. Automatic Morphological Analysis of Basque. Literary and Linguistic Computing 11 (4): 193-203. Oxford University Press. 1996.

(Aduriz *et al.*, 96) Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. EUSLEM: A lemmatiser/tagger for Basque. *Proc. of the Euralex´96*, Goteborg, Sweden. 1996.

(Ananiadou, 94) Ananiadou S. A Methodology for Automatic Term Recognition. *Proc. of the Conference on Computational Linguistics (Coling-94)*, 1034-1038, Kyoto, Japan. 1994.

(Arpper, 95) Arpper A. Term Extraction from Unrestricted Text. *http://www.lingsoft.fi/doc/nptool/term-extraction.html*.

(Bourigault, 92) Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases. *Proc. of the Conference on Computational Linguistics (Coling-92)*, 977-981, Nantes, France. 1992.

(Bourigault & Habert, 98) Bourigault D., Habert B. Evaluation of Terminology Extractors: Principles and Experiments. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),299-305.* Granada,1998.

(Brill, 92) Brill E. A simple rule-based part of speech tagger. ANLP'92, 152-155.

(Church & Hanks, 90) Church K., Hanks P. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics* **16**:22-29. 1990.

(Church & Dagan, 94) Church K., Dagan I. Termight: Identifying and Traslating Technical Terminology. *Proc. of the 4th Conference on Applied Language Processing*, Stuttgart, Germany. 1994.

(Cutting *et al.*, 92) Cutting D., Kupiec J., Pedersen J., Sibun P. A practical part-of-speech tagger. ANLP'92, 133-140.

(Daille & Jacquemin, 98) Daille B., Jacquemin C. Lexical Database and Information Access: A Fruitful Association?. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),669-673.* Granada,1998.

(El-Hadi & Jouis, 98) El-Hadi W.M., Jouis C. Terminology Extraction and Acquisition from Textual Data: Criteria for Evaluating Tools and Methods. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),1.175-1.178.* Granada,1998.

(Ezeiza *et al.*, 98) Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. Combining stochastic and rule-based methods for desanbiguation in agglutinative languages. *Proc. of the Conference on Computational Linguistics (Coling-ACL'98)*. Montreal 1998.

(Ferrari & Prince, 96) Ferrari K.T., PrinceS. Création et Extension Automatiques de Dictionnaries Terminologiques Multilingues Specialisés a partir de Corpus Monolingues. *Proc. of the Conference on Natural Language Processing and Industrial Applications,* 79-86. Moncton, N.B. Canada. 1996.

(Frantzi & Ananiadou, 96) Frantzi K.T., Ananiadou S. Extracting Nested Collocations. *Proc. of the Conference on Computational Linguistics (Coling-96)*, 41-46. 1996.

(Hamon & Nazarenko, 98) Hamon T. & Nazarenko A. Using a general semantic information to help terminology structuration. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),675-680.* Granada,1998.

(Justeson & Katz, 95) Justeson J.S., Katz S.M. Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* **1** (1): 9-27. Cambridge University Press. 1995.

(Maynard & Ananiadou, 98) Maynard D. & Ananiadou S. Term Sense Disambiguation using Domain-Specific Thesaurus. *Proc. Of the First Conference on Language Resources & Evaluation (LREC'98),681-685.* Granada,1998.

(Oueslati *et al.*, 96) Oueslati R., Frath P., Rousselot F. Term Identification and Knowledge Extraction. *Proc. of the Natural Language Processing and Industrial Applications (NLP+IA)*, 191-196, Moncton, Canada. 1996.

(Su *et al.*, 96) Su K., Wu M., Chang J. A Corpus-based Aproach to Automatic Compound Extraction. *Proc. of the Conference on Computational Linguistics (Coling-96)*, 243-247. 1996.

Resumen

En los últimos años se vienen desarrollando herramientas para la extracción automática de terminología, principalmente para el inglés pero también para otras lenguas. Los resultados son prometedores. Como ejemplo de ello podemos nombrar muchas herramientas, entre ellas LEXTER, Termight de AT&T, TERMS de IBM y NPtool.

Las áreas de aplicación de estas herramientas se pueden dividir en dos grandes grupos: trabajo terminológico propiamente dicho e indexación de información. Para la construcción de estas herramientas se suelen combinar técnicas estadísticas con técnicas lingüísticas, aunque en muchos casos se da prioridad casi exclusiva a las primeras (Su et al., 96). Sin embargo, cuando la lengua de trabajo es de flexión rica, o todavía más, si es aglutinante como el vasco, el proceso lingüístico se vuelve imprescindible.

El proceso lingüístico básico es el etiquetado morfosintáctico, ya que además de ser fundamental para buscar los patrones correspondientes a la estructura de los términos, puede ser un preproceso para un análisis sintáctico superficial.

En nuestro sistema usamos un etiquetador/lematizador (Ezeiza et al., 98) que combina técnicas estadísticas y conocimiento lingüístico, y que además de asignar las etiquetas desambigua el lema, pudiéndose así utilizar medidas estadísticas en función de los lemas y no sólo de las formas.

Aunque el proyecto esta todavía en fase de desarrollo y no tenemos resultados, en el artículo se hará, además de una amplia revisión bibliográfica, una descripción de las partes ya en funcionamiento así como de la arquitectura del sistema completo y de los trabajos en curso.