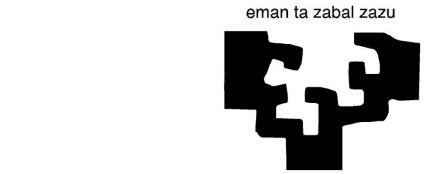
Euskal Herriko Unibertsitatea / Universidad del País Vasco



Departamento de Lenguajes y Sistemas Informáticos

Entidades Nombradas en Euskera: Identificación, Clasificación, Traducción y Desambiguación

Izaskun Fernandez Gonzalez

Tesis presentada para obtener el título de

Doctora en Informática

Donostia, Febrero del 2012.

CAPÍTULO I

Introducción

Los humanos somos capaces de leer y entender fácilmente la información que recibimos y que día a día va en aumento. Sin embargo, hoy en día los computadores no son aún capaces de ello y están muy lejos de poder entender toda esa información, al menos cuando se trata de información no estructurada como la representada en forma de texto: contenidos publicados en páginas web o blogs; noticias y/o conversaciones de las redes sociales; y en general, todo tipo de contenido digital accesible en la web. De todas formas, y a pesar de que la interpretación y la comprensión de dicha información aún no es un hecho real, las investigaciones en el ámbito del procesamiento del lenguaje natural hacen que esa realidad esté cada vez más cerca.

De hecho, en el camino hacia la interpretación y comprensión automática, ya se han definido ciertos problemas básicos para llegar a esa comprensión que se han resuelto con procesos automáticos. Ejemplo de esos problemas básicos ya resueltos para varios idiomas son el análisis morfológico y el sintáctico. Gracias a los procesos que automatizan ambos análisis es posible extraer información básica de los textos de forma automática. Información muy valiosa para el desarrollo de aplicaciones de mayor alcance como pueden ser aplicaciones de traducción automática y/o sistemas de pregunta-respuesta.

Dentro de esos procesos básicos del ámbito del procesamiento del lenguaje natural precisamente se encuentra el objetivo principal de esta tesis, el tratamiento automático de las entidades nombradas en euskera. Se entiendo como entidades nombradas aquellas expresiones que hacen referencia a nombres

propios de personas, lugares y organizaciones[1].

Una de las primeras tareas o problemas a resolver a la hora de trabajar con entidades nombradas es la identificación y clasificación de estas en textos escritos. Es decir, dada la frase "*Izaskun está realizando sus estudios de doctorado en la Universidad del País Vasco*", trata de identificar las expresiones *Izaskun* y *Universidad del País Vasco* como entidades y a continuación las clasifica como persona y organización respectivamente.

A pesar de que la identificación tiene sus propias dificultades, la clasificación de las entidades nombradas es la tarea más compleja de las dos, ya que de haber algún tipo de ambigüedad en la expresión, la clasificación es la tarea que deberá de resolver dicho problema, y no la identificación.

En la clasificación de las entidades nombradas, la ambigüedad puede darse de dos formas: por un lado, la misma expresión puede tener dos clasificaciones diferentes, es decir, la misma expresión puede referenciar a dos tipos de entidades nombradas diferentes; y por otro lado, a pesar de pertenecer a la misma categoría, una expresión puede hacer referencia a dos entidades nombradas diferentes. Ejemplo de los dos casos de ambigüedad son los que se mencionan a continuación:

- *Walt Disney*: esta expresión puede hacer referencia tanto a la compañía de dibujos animados como a la persona que la fundó, pudiendo así la misma expresión hacer referencia a una entidad nombrada tanto de tipo persona como de tipo organización.
- *Armstrong*: con esta expresión se puede hacer referencia tanto al ciclista *Lance Armstrong* como al astronauta *Neil Armstrong*. Por tanto con ese apellido se pueden referenciar dos entidades nombradas diferentes de la misma categoría (persona).

El primer tipo de ambigüedad influye directamente en el problema de la clasificación, en cambio el segundo no, ya que a pesar de la ambigüedad, la categoría a asignar es la misma y, por tanto, no supone ningún problema añadido para la tarea de clasificación.

Sin embargo, no resolver el segundo tipo de ambigüedad, sí puede perjudicar a un sistema de mayor alcance, como por ejemplo un sistema de pregunta-respuesta, dado que para un sistema de este tipo es necesario saber

[1] Las expresiones temporales, númericas, etc. quedan fuera del alcance de esta tesis, ya que se resuelven en un proceso previo a los aquí descritos.

a cuál de los dos Armstrong se está haciendo referencia, al ciclista o al astronauta, siendo esta información necesaria para una respuesta coherente. Por tanto, la desambiguación de las entidades nombradas también es necesaria para poder crear aplicaciones robustas de ese tipo.

Además si nos situamos en un entorno multilingüe en el que las entidades nombradas pueden aparecer escritas en diferentes idiomas, las tareas de identificación, clasificación y desambiguación de dichas expresiones no son suficientes para poder relacionar las expresiones en diferentes idiomas que hagan referencia a la misma entidad. Para dicho propósito es necesario conocer la forma de la entidad en cada uno de los idiomas. Una de las vías posibles para obtener esa información es la tarea de traducción de entidades, tarea que tiene como objetivo traducir una entidad nombrada en un idioma de partida a un idioma destino.

Si en lugar de traducir las entidades nombradas de forma particular, se traducen con las estrategias más genéricas de los sistemas de traducción automática, no nos debe sorprender encontrar traducciones del tipo *school of the right of Harvard* para la entidad en castellano *Escuela de Derecho de Harvard*, ya que las entidades nombradas y el conjunto de elementos que las forman muchas veces no comparten el mismo comportamiento que el resto de los elementos dentro de un esquema de traducción, y es por eso que se traducen mal. Por tanto es necesario definir un tratamiento, aparte de la estrategia general de traducción que se encargue exclusivamente de automatizar la traducción de las entidades nombradas, y mediante ambas estrategias poder definir así un sistema de traducción automática robusto y completo.

El uso de estrategias que solventen la identificación, clasificación, traducción y desambiguación automática de las entidades nombradas, por tanto, puede ayudar a mejorar el comportamiento de las aplicaciones de mayor alcance mencionadas al principio del capítulo. Un ejemplo claro pueden ser los ya mencionados sistemas de pregunta-respuesta, en los que cada una de las tareas descritas pueden ayudar a reforzar diferentes aspectos del sistema:

- Tener identificadas y clasificadas las entidades nombradas tanto en el conjunto de preguntas como en el de respuestas puede ser muy útil para reducir el tamaño del conjunto de documentos sobre los que buscar la respuesta correcta, ayudando así a optimizar el proceso de búsqueda.
- En el caso que en el conjunto de respuestas pueda haber documentos escritos en diferentes idiomas, y por tanto poder conocer las formas

en diferentes idiomas de la entidad nombrada sobre la que se pregunta puede ayudar al sistema a dar una respuesta multilingüe más completa.

- Finalmente, tener resuelta la ambigüedad de las apariciones ambiguas de las entidades nombradas tanto en la colección de preguntas como de respuestas puede resultar beneficioso para que el sistema pueda dar una respuesta más precisa, evitando devolver información acerca de otras acepciones a las que se puedan hacer referencia con la misma expresión.

I.1. Objetivos

El objetivo principal, por tanto, de esta tesis es automatizar el tratamiento de las entidades nombradas en euskera. Para lograr dicho objetivo, se han establecido tres criterios metodológicos:

1. siendo el euskera un idioma de escasos recursos, se ha priorizado la reutilización de recursos y el uso de métodos no supervisados y semi-supervisados.

2. trabajar tanto con técnicas basadas en el conocimiento de un idioma como con técnicas de aprendizaje automático, combinando ambas cuando esto sea posible. Se trata así de evitar el uso de técnicas muy sofisticadas, apostando por la combinación de métodos simples y pequeñas modificaciones en estos cuando sea necesario.

3. analizar el impacto de las características morfosintácticas propias del euskera al tratar de automatizar el tratamiento de las entidades nombradas.

Siguiendo estos tres criterios y dentro del objetivo principal de esta tesis, se pretenden abordar tres tareas principales:

1. *Identificación y clasificación de entidades nombradas en euskera.*

   El objetivo principal de esta tarea es el desarrollo de una herramienta que sea capaz de identificar y clasificar automáticamente las entidades nombradas en textos escritos en euskera, de precisión semejante a las desarrolladas para el inglés. Para ello, se analizarán y aplicarán técnicas tanto basadas en el conocimiento como basadas en aprendizaje

automático, así como la combinación de ambas, con el fin de analizar el comportamiento de las combinaciones y obtener un sistema robusto. En el caso del euskera, al tener que trabajar con recursos limitados se tratará de hacer frente a los problemas que dicha limitación genere.

2. *Traducción de entidades nombradas en euskera.*

   La generación automática de referencias multilingües de entidades nombradas, que pueden resultar muy útiles tanto en aplicaciones de traducción automática como en sistemas de pregunta-respuesta multilingüe, es el próposito principal de esta tarea. Se pretenden realizar diversas aproximaciones para analizar y evaluar los comportamientos de diferentes técnicas, algunas basadas en el conocimiento del idioma y otras semisupervisadas, que sirvan para abordar el problema de la traducción de las entidades nombradas, siendo el euskera el idioma de partida y el castellano el idioma destino. Se estudiarán los recursos necesarios para cada aproximación y se analizarán los resultados de las diferentes técnicas tratando de realizar una comparativa de estas. Finalmente, se estudiará la portabilidad de la aproximación basada en técnicas semisupervisadas a otros idiomas diferentes al euskera-castellano.

3. *Desambiguación de entidades nombradas en euskera.*

   El problema que resuelve esta tarea es la desambiguación automática de las apariciones ambiguas de las entidades nombradas en los textos escritos en euskera. Como en cualquier tarea de desambiguación, para poder resolver automáticamente la ambig
   üedad, además del contexto de la aparición ambigua, es necesaria una base de conocimiento en la que se describan las posibles acepciones de las expresiones ambiguas. Para cubrir esta necesidad se pretende analizar dentro de esta tarea la validez de la Wikipedia en euskera para la generación de dicho repositorio. De esta forma será posible la definición de un proceso que sea capaz de relacionar una aparición ambigua de una entidad nombrada en euskera con su correspondiente entrada de la Wikipedia. Para la automatización de dicho proceso se tratará de aplicar aquellas técnicas con las que se hayan conseguido los mejores resultados para otros idiomas, como por ejemplo el inglés, y

analizar su comportamiento al utilizarlas con recursos limitados como es el caso del euskera.

Por tanto, en esta tesis además de marcarnos el objetivo principal de desarrollar herramientas para la identificación, clasificación, traducción y desambiguación automática de entidades nombradas en euskera, es nuestro propósito también estudiar y comparar el comportamiento de diferentes estrategias en entornos de recursos limitados. Para que dichas comparaciones sean posibles, en cada tarea se han utilizado diferentes técnicas, con el objetivo de que, además de servir para probar su validez para el euskera, sirvan estos estudios de recursos y técnicas para otros idiomas de características similares.

I.2. Estructura del documento

En el resto de capítulos de este documento se describen los trabajos realizados para resolver cada una de las tareas identificadas en este primer capítulo en torno a las entidades nombradas en euskera. Hemos dedicado un capítulo a cada tarea, por tanto, este documento consta de tres capítulos centrales.

En el capítulo II se describen los experimentos realizados para la identificación y clasificación de entidades nombradas en euskera. En el siguiente capítulo se detallan las diferentes aproximaciones realizadas para traducir entidades nombradas en euskera al castellano. Y en el capítulo IV, el último de los tres centrales, se detallan los experimentos llevados a cabo para tratar de relacionar apariciones de entidades nombradas ambiguas en euskera con entradas de la Wikipedia en euskera.

Para finalizar, en el capítulo V se presentan las conclusiones y aportaciones surgidas de esta tesis, así como las líneas que quedan abiertas para futuros trabajos de investigación que den continuidad a los trabajos aquí presentados.

Además de los capítulos mencionados, este documento consta de cuatro anexos: el anexo A describe la gramática para la identificación de entidades nombradas en euskera, que se ha definido para el desarrollo de la herramienta basada en el conocimiento del euskera; en el anexo B se detalla el sistema de categorías de EDBL que se ha utilizado a lo largo de toda la tesis; en el anexo C se describe la gramática que reúne las reglas de transformación fonológica para traducir las entidades nombradas en euskera al castellano; y, finalmente,

en el anexo D se detallan las reglas de ordenación de los elementos a aplicar al traducir las entidades nombradas en euskera al castellano.

I.3. Publicaciones

Para concluir con este capítulo introductorio se listan a continuación los artículos publicados a lo largo de tesis. Por un lado, se listan los estrechamente ligados a la tesis; y por otro lado, se mencionan algunos artículos menos relacionados con la tesis, pero que sí están realizados en el ámbito general de la tesis.

I.3.1. Las publicaciones estrechamente ligadas a la tesis

Las publicaciones acerca la identificación y clasificación de las entidades nombradas en euskera, descritas en el capítulo II son:

- I. Alegria, N. Ezeiza, I. Fernandez, R. Urizar. Named Entity Recognition and Classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI2003*, 2003.
- I. Alegria, O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, R. Urizar. Design and development of a named entity recognizer for an agglutinative language. *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*, 2004.
- I. Alegria, O. Arregi, N. Ezeiza, I. Fernandez. Lessons from the Development of a Named Entity Recognizer. *Procesamiento del Lenguaje Natural, vol. 36, p. 25-37*, 2006.

Las publicaciones acerca la traducción de las entidades nombradas en euskera, descritas en el capítulo III son:

- I. Alegria, N. Ezeiza, I. Fernandez. Named Entities Translation Based on Comparable Corpora. *MultiWord Expressions in a Multilingual Context Workshop on EACL06, p.1-8*, 2006.
- I. Alegria, N. Ezeiza, I. Fernandez. Translating Named Entities using Comparable Corpora. *Building and Using Comparable Corpora. LREC 2008 Workshop*, 2008.

- I. Fernandez, I. Alegria, N. Ezeiza. Using Wikipedia for Named Entities Translation. *SALTMIL2009 workshop: IR-IE-LRL*, 2009.

Finalmente, la publicación acerca la desambiguación de las entidades nombradas en euskera, descrita en el capítulo IV es:

- I. Fernandez, I. Alegria, N. Ezeiza. Semantic Relatedness for Named Entity Disambiguation using a Small Wikipedia. *TSD 2011, LNAI 6836, p. 276-283*, 2011.

I.3.2. Otras publicaciones

- O. Ansa, X. Arregi, B. Arrieta, A. Díaz de Ilarraza, N. Ezeiza, I. Fernandez, A. Garmendia, K. Gojenola, B. Laskurain, E. Martínez, M. Oronoz, A. Otegi, K. Sarasola, L. Uria. Integrating NLP Tools for Basque in Text Editors. *Workshop on International Proofing Tools and Language Technologies*, 2004.
- A. Jimenez, I. Fernandez, D. Pérez, E. Viejo, F.J. Díez, de X. G. Kortazar, M. García, V. Maojo, A. Cobo, F. del Pozo. Patient-based Literature Retrieval and Integration - A Use Case for Diabetes and Arterial Hypertension. *In Proceedings of HEALTHINF 2011*, 2011.
- I. Fernandez, A. Jimenez, X. G. Kortazar, D. Perez. A New Method to Retrieve, Cluster And Annotate Clinical Literature Related To Electronic Health Records. *Proceedings of the Third International Workshop on Health Document Text Mining and Information Analysis (LOUHI)*, 2011.

CAPÍTULO II

Identificación y Clasificación de Entidades Nombradas

II.1. Resumen

El reconocimiento de entidades nombradas (NERC), tal y como se definió en la conferencia *Message Understanding Conference* (MUC) (Chinchor, 1998), trata de extraer de textos escritos entidades nombradas como las expresiones que definen los nombres de personas, lugares y organizaciones, así como expresiones temporales y numéricas. A menudo esta tarea se divide en dos subtareas: una subtarea de identificación que trata de identificar los elementos de las entidades nombradas (NE) y una segunda de clasificación (NEC) que se encarga de clasificar las expresiones identificadas en la primera. Siguiendo esta división de tareas es como hemos tratado de resolver la tarea de NERC en textos escritos en euskera.

Para abordar la subtarea tanto de identificación como de clasificación hemos aplicado técnicas tanto basadas en el conocimiento del idioma y basadas en aprendizaje automático, así como la combinación de ambas, tal y como se describe en los trabajos de Alegria *et al.* (2003) y Alegria *et al.* (2004) y Alegria *et al.* (2006a) respectivamente.

Como se puede ver en la publicación de Alegria *et al.* (2003), nuestra primera aproximación para abordar el problema de identificación y clasificación de NE es una aproximación lingüística basada en el conocimiento del

euskera, ya que tanto para la fase de identificación como para la de clasificación se han definido gramáticas y heurísticos basados en sus características lingüísticas, que se han obtenido tras una revisión exhaustiva manual por parte de expertos lingüistas. Los resultados de esta herramienta lingüística (*Eihera*), a pesar no estar demasiado cerca de los resultados de herramientas de otros idiomas como el inglés, *Eihera* ha sido una herramienta muy útil, por un lado para crear corpora etiquetados semiautomáticamente, y por otro para la identificación de atributos relevantes a utilizar en los experimentos realizados con técnicas de aprendizaje automático supervisado tal y como se muestra en las otras dos publicaciones.

Otro aspecto que se presenta en las dos últimas publicaciones de este problema es la aproximación de la combinación de diferentes técnicas, para conseguir sistemas más robustos. En (Alegria *et al.*, 2004) se presentan, por un lado, los resultados de la fase de identificación resuelto con diferentes algoritmos de aprendizaje automático y, por otro lado, los resultados obtenidos combinando diferentes técnicas.

En (Alegria *et al.*, 2006a) en cambio, no sólo se presentan los resultados obtenidos para la tarea de clasificación sino que se presentan también los resultados de un sistema NERC para el euskera completo, abordando tanto la identificación como la clasificación. En ambos trabajos se refleja lo beneficioso que resulta la combinación de sistemas basados en diferentes técnicas, consiguiendo mejoras importantes en los resultados.

De todos los experimentos realizados uno de los sistemas más robustos para la tarea NERC en euskera es el sistema basado en el algoritmo de aprendizaje automático AdaBoost, concretamente la implementación descrita en el trabajo (Carreras *et al.*, 2003), llamado *Abionet*. Este sistema fue el ganador en la edición de 2002 CoNLL para el castellano y el holandés, y en el caso del euskera, además de ser uno de los más robustos, forma parte de la mejor combinación que se ha conseguido en este trabajo. Los resultados de *Abionet* para el euskera, tanto evaluado independientemente como en combinación con otros sistemas están detallados en la última fila de la Tabla II.1 (la equivalente a la Tabla II.24 de la versión en euskera). Junto con los resultados para el euskera, se presentan los resultados de *Abionet* para los diferentes idiomas de las ediciones 2002 y 2003 en el marco de la tarea compartida de CoNLL.

A pesar de no conseguir igualar los resultados de herramientas de idiomas tan trabajados como el inglés, podemos decir que hemos conseguido desarrollar una herramienta (*Eihera+*), basada en la combinación de diferen-

	Abionet	*Combinación*
Inglés	85,00 %	90,30 %
Castellano	81,39 %	-
Holandés	77,05 %	-
Alemán	69,15 %	74,17 %
Euskera	65,24 %	**71,35 %**

Tabla II.1: Evaluación de la herramienta *Abionet* (F_1) para diferentes idiomas

tes estrategias, cercana a los resultados de otros idiomas como por ejemplo el Alemán, de recursos más limitados y de morfología rica.

Además de los resultados presentados en los artículos relacionados a este capítulo, en la memoria en euskera hemos realizado un repaso bibliográfico extenso y presentamos una evaluación detallada por categorías en la tarea de clasificación, así como un estudio de selección de atributos e instancias especialmente en los experimentos basados en técnicas de aprendizaje automático. Se ha tenido muy en cuenta en ese estudio tanto la información interna como externa de las NEs. Se entiende como interna la información que los propios elementos de las NEs pueden aportar, y como externa la información proveniente de fuentes externas, como pueden ser listas especializadas de NEs (conocinadas como *gazetteer*s) o listas de elementos fuera de las NEs que pueden aportar información sobre ellas.

II.2. Publicaciones

A continuación se adjuntan las publicaciones relacionadas con este capítulo:

- I. Alegria, N. Ezeiza, I. Fernandez, R. Urizar. Named Entity Recognition and Classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI2003*, 2003.
- I. Alegria, O. Arregi, I. Balza, N. Ezeiza, I. Fernandez, R. Urizar. Design and development of a named entity recognizer for an agglutinative language. *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*, 2004.
- I. Alegria, O. Arregi, N. Ezeiza, I. Fernandez. Lessons from the Development of a Named Entity Recognizer. *Procesamiento del Lenguaje*

Natural, vol. 36, p. 25-37, 2006.

CAPÍTULO III

Traducción de Entidades Nombradas

III.1. Resumen

En este capítulo hemos abordado la problemática de la generación automática de referencias multilingües de entidades nombradas. Hemos realizado un amplio repaso bibliográfico de los trabajos existentes en el ámbito de la traducción de este tipo de expresiones y hemos planteado diferentes aproximaciones que resuelven esta temática: una aproximación lingüística basada en el comportamiento del euskera y el castellano, y otras semisupervisadas que tratan de resolver el problema de la forma más independientemente posible a los idiomas que se abordan. A pesar de seguir diferentes estrategias, en ambos casos el recurso clave ha sido un corpus comparable.

En el primer trabajo publicado sobre el tema de este capítulo (Alegria *et al.*, 2006b) se describen los sistemas y los experimentos realizados con ambas aproximaciones para traducir NEs de euskera a castellano. En este trabajo se puede ver como el sistema basado en técnicas semisupervisadas e implementado con una metodología semiindependiente de los idiomas, a pesar de no igualar los resultados del sistema lingüístico desarrollado e implementado con información morfosintáctica exacta del euskera y castellano, consigue unos resultados prometedores, en términos de F_1 sólo 2,3 puntos por debajo del lingüístico.

La evaluación de portabilidad de esta solución semiindependiente de los idiomas a traducir se describe en el trabajo (Alegria *et al.*, 2008). Concreta-

mente se presenta la construcción y evaluación de la herramienta tanto para la traducción de NEs de euskera a castellano como de castellano a inglés, reportando considerablemente peores resultados para el último par de idiomas. Sin embargo, y a pesar de estos resultados, no podemos decir que la solución no sea portable a otros idiomas de forma satisfactoria, ya que, tal y como se describe en la publicación, al realizar un análisis de los errores ocurridos se ha visto que la calidad del corpus comparable es de vital importancia en esta solución, y precisamente esa calidad no equivalente en los corpora de este par de idiomas. En cualquier caso, lo que sí queda reflejado en la publicación es la facilidad de construir un sistema para un nuevo par de idiomas siguiendo la estrategia del sistema basado en técnicas semisupervisadas.

En la última publicación ligada a este capítulo (Fernandez *et al.*, 2009), se presenta una aproximación basada en la solución semiindependiente de los idiomas, pero además de seguir la estrategia de dicha solución se describe la incorporación de algunas nuevas fases que explotan las características de la Wikipedia, como son, por ejemplo, los links entre entradas de Wikipedias de diferentes idiomas (denominados WIL) que definen equivalencias entre dichas entradas. En este último trabajo, por tanto, se describe un sistema de traducción basado en la Wikipedia, que se evalúa en el escenario de traducción de NEs en euskera al inglés. Además de los resultados del artículo, hemos realizado una evaluación más detallada en la que se describen el número de aciertos en cada fase y la aportación de la inclusión de las características de la Wikipedia en el sistema. Estos resultados son los que se detallan en la Tabla III.1 (la equivalente a la Tabla III.11 de la memoria en euskera).

Fase	**Nº total de NEs**	**Nº de traducciones correctas**
Diccionario	10	10
Wikipedia[1]	366	362
Proceso de traducción[2]	49	46
Sin traducción	75	0

[1] *Traducciones obtenidas buscando directamente la forma de euskera de la NE en la Wikipedia en inglés*

[2] *Traducciones de NEs obtenidas tras traducir uno a uno todos sus elementos, construyendo candidatos enteros con las traducciones parciales de los elementos y finalmente aplicando la fase de selección de el(los) candidato(s) más adecuado(s)*

Tabla III.1: Dsitribución de traducciones

Además en la memoria en euskera se describe en detalle cada uno de los sistemas arriba mencionados, así como las principales características y recursos que los hacen diferentes. Se describen en detalle también el origen de todos los corpora de evaluación que se han utilizado a lo largo de los experimentos y en qué evaluaciones se han utilizado. Gracias a esa correspondencia entre corpora de evaluación y sistemas, se presenta una clara comparación de sistemas cuando los corpora lo permiten. Finalmente, se define también la metodología de construcción automática de corpora comparables basada en una versión etiquetada de la Wikipedia. Gracias a esta metodología y a la aproximación semiindependiente de los idiomas, contando con un diccionario bilingüe para un nuevo par de idiomas, hemos conseguido una solución que nos permite crear un sistema de traducción de NEs con muy poco esfuerzo.

III.2. Publicaciones

A continuación se adjuntan las publicaciones relacionadas con este capítulo:

- I. Alegria, N. Ezeiza, I. Fernandez. Named Entities Translation Based on Comparable Corpora. *MultiWord Expressions in a Multilingual Context Workshop on EACL06, p.1-8*, 2006.
- I. Alegria, N. Ezeiza, I. Fernandez. Translating Named Entities using Comparable Corpora. *Building and Using Comparable Corpora. LREC 2008 Workshop*, 2008.
- I. Fernandez, I. Alegria, N. Ezeiza. Using Wikipedia for Named Entities Translation. *SALTMIL2009 workshop: IR-IE-LRL*, 2009.

CAPÍTULO IV

Desambiguación de Entidades Nombradas

IV.1. Resumen

La desambiguación automática de las apariciones ambiguas de las entidades nombradas en los textos escritos en euskera es el problema que hemos tratado de resolver en este capítulo. Y con el fin de identificar las mejores estrategias de desambiguación de NEs hemos realizado un amplio repaso bibliográfico de los trabajos existentes. Dado que es una tarea relativamente reciente, y a pesar de que se han analizado algunos trabajos previos, la mayor parte de este repaso lo hemos centrado en las soluciones que se han presentado en las tareas compartidas de Enlace de Entidades (*Entity Linking*) en el marco de las ediciones del 2009 y 2010 de la tarea compartida TAC-KBP.

Siguiendo el criterio de la mayoría de trabajos existentes, hemos diseñado una solución que entiende el problema de desambiguación como un problema que trata de relacionar una aparición ambigua de una NE con una entrada de una base de conocimiento. En nuestro trabajo, siguiendo de nuevo la estrategia de TAC-KBP, hemos utilizado la Wikipedia de euskera como base de conocimiento. Para la definición del diccionario que describe la correspondencia entre una aparición ambigua y sus posibles acepciones en la base de conocimiento, es decir, sus posibles entradas de la Wikipedia, hemos seguido los pasos descritos en el trabajo de Agirre *et al.* (2009). Para la selección de la entrada más adecuada, hemos aplicado diferentes algoritmos. Todos ellos y sus evaluaciones se detallan en la publicación de Fernandez *et al.* (2011).

Como se puede ver en dicha publicación la primera fase es común a todos los sistemas, lo que implica que la única forma de conseguir las acepciones de una aparición ambigua de una NE es consultando el diccionario de mapeo previamente mencionado. Por tanto la variación de los resultados depende de la calidad con la que el algoritmo de desambiguación resuelve el problema de selección de la mejor acepción. Al hilo de esa calidad de desambiguación se describe en el artículo una propuesta de mejora para uno de los mejores algoritmos de desambiguación de NEs en inglés, el algoritmo ESA (*Explicit Semantic Analysis*), tratando de minimizar la influencia negativa que parece tener el uso de Wikipedias más pequeñas y descompensadas, como es el caso de la del euskera, sobre este algoritmo. A lo largo del artículo, y en otras evaluaciones presentadas en la memoria, se confirma que la propuesta de mejora que planteamos es acertada, ya que consigue en todas las evaluaciones mejorar los resultados de ESA para el euskera. Para mayor detalle se puede consultar el artículo de Fernandez *et al.* (2011) donde aparece descrito el factor de corrección que proponemos como mejora.

Además de los experimentos descritos en ese artículo, en la memoria en euskera, se presentan dos evaluaciones más: una que corresponde a los mismos datos descritos en el artículo, pero en lugar de utilizar todos los casos de los corpora de evaluación únicamente, utilizando aquellos ejemplos en los que, tras la consulta al diccionario la aparición ambigua se puede referir a más de una entrada de la Wikipedia, eliminando de la evaluación el resto de casos; y por otro una evaluación utilizando datos lematizados.

En el primer caso, eliminando del corpus de evaluación aquellos ejemplos que según el diccionario de mapeo únicamente se pueden relacionar con una entrada de la Wikipedia, sin evaluar si es o no la acepción correcta[1], como era de esperar hemos comprobado que los resultados empeoran notablemente respecto a la evaluación realizada con el corpus completo, tal y como se puede ver en la Tabla IV.1 (la equivalente a la Tabla IV.3 en la memoria en euskera), pero las diferencias entre sistemas se reducen.

En la evaluación de los sistemas con datos lematizados ocurre algo similar. Los resultados en todos los sistemas empeoran notablemente respecto a los obtenidos al evaluarlos con datos únicamente tokenizados, tal y como se

[1]Por el hecho de que exista una única entrada en la Wikipedia que pueda ser la NE que desambigüe una aparición, no tiene porqué ser la adecuada, ya que pueden existir apariciones ambiguas que no tenga la acepción correcta descrita en la Wikipedia. En cualquier caso, en este trabajo no se ha realizado dicho análisis, ni en este proceso ni en ninguno de los sistemas descritos.

	Precision	Recall	F_1
MFS - sólo ambiguos	53,00 %	53,00 %	53,00 %
VSM - sólo ambiguos	62,38 %	52,62 %	57,10 %
ESA - sólo ambiguos	57,75 %	47,40 %	52,00 %
bESA - sólo ambiguos	65,51 %	56,10 %	60,44 %
UKB - sólo ambiguos	74,50 %	71,76 %	**73,10 %**

Tabla IV.1: Evaluación sobre el corpus A tras la eliminación de ejemplos no ambiguos

refleja en la Tabla IV.2 (la equivalente a la Tabla IV.8 de la memoria en euskera).

	MFS	VSM	ESA	bESA
***A*-Tokenizado**	68,32 %	75,53 %	72,43 %	77,66 %
***A*-Lematizado**	68,32 %	66,04 %	58,58 %	73,96 %
$B_{Desarrollo}$-Tokenizado	72,00 %	70,48 %	61,09 %	68,00 %
$B_{Desarrollo}$-Lematizado	72,00 %	67,10 %	55,26 %	**70,86 %**
B_{Test}-Tokenizado	70,40 %	70,00 %	61,60 %	68,40 %
B_{Test}-Lematizado	70,04 %	69,00 %	54,40 %	**68,50 %**

Tabla IV.2: Evaluaciones de los sistemas con datos únicamente tokenizados y lematizados (resolviendo los empates con la estrategia de MFS)

En la memoria en euskera se describe también el trabajo de revisión de errores que se ha realizado. En dicha revisión se han analizado los errores que se producen en el proceso de desambiguación aplicando diferentes algoritmos tratando a la vez de identificar los puntos fuertes y débiles de cada uno de ellos para tratar de encontrar posibles mejoras.

IV.2. Publicaciones

A continuación se adjunta la publicación relacionada con este capítulo:

- I. Fernandez, I. Alegria, N. Ezeiza. Semantic Relatedness for Named Entity Disambiguation using a Small Wikipedia. *TSD 2011, LNAI 6836, p. 276-283*, 2011.

CAPÍTULO V

Conclusiones y trabajos futuros

En este capítulo se describen por un lado las aportaciones del trabajo realizado en esta tesis, y por otro lado las conclusiones que hemos obtenido de las mismas. Para finalizar y ligado al trabajo realizado, se presentan las posibles líneas futuras de trabajo.

V.1. Aportaciones

A lo largo del documento se han presentado los trabajos realizados en el área de las entidades nombradas (NE) en euskera. En concreto, se han abordado todos los objetivos fijados en el capítulo I, como son la identificación y clasificación (NERC), la traducción (NET) y la desambiguación (NED) de las entidades nombradas en euskera. En la elaboración de todos ellos, y siguiendo los criterios metodológicos establecidos en la misma fase de defición de objetivos, se ha trabajado en soluciones basadas en la reutilización de recursos, así como en métodos simples y combinados, evitando el uso de técnicas muy sofisticadas. En todas ellas también se ha analizado en detalle el impacto que suponen las características morfológicas del euskera.

Siendo el tratamiento de las entidades nombradas un área muy extensa, ha resultado imposible profundizar en todos los campos. Por ese motivo, en este trabajo hemos tratado de acotar los problemas de identificación, clasificación, traducción y desambiguación de NE a escenarios reales adecuados para un idioma de recursos limitados. En esos escenarios es donde hemos

abordado cada uno de los problemas, y hemos tratado de buscar soluciones. Son precisamente las tres herramientas que resuelven esos problemas las aportaciones más relevantes de este trabajo: la herramienta de identificación y clasificación de NE en euskera; el traductor automático de NEs; y finalmente la herramienta de desambiguación que relaciona apariciones de NE en euskera con la Wikipedia.

Todas esas herramientas se han desarrollado con recursos límitados, siendo la reutilización un requisito que se ha tenido muy en cuenta. La reutilización se ha considerado tanto desde el punto de vista de los recursos como de los diseños. Por un lado, hemos tratado de explotar y utilizar todos los recursos disponibles. Y por otro lado, hemos pensado en diseños que en la medida de lo posible sean reutilizables para otros idiomas de recursos limitados como el euskera.

Respecto a los recursos utilizados, hay que mencionar el ánalisis y explotación que hemos realizado de recursos bien conocidos como son WordNet y la Wikipedia, tratando de utilizar su información en la mayor medida posible resolver los objetivos planteados al inicio de la tesis. En el caso de la Wikipedia, podemos incluso considerarnos pioneros ya que a pesar de ser hoy en día un recurso de uso muy extendido, al inicio de nuestro trabajo en torno a la traducción de NEs, al menos en lo que al área del procesamiento del lenguaje natural respecta, fuimos de los primeros en explotar esta enciclopedia.

En cuanto a las estrategias empleadas para el diseño e implementación de las herramientas, hemos aplicado las dos aproximaciones principales del procesamiento del lenguaje natural: la basada en el conocimiento de la lengua y la basada en aprendizaje automático. El hecho de utilizar las dos aproximaciones nos ha permitido comparar los comportamientos de ambas en las tareas abordadas. Además siempre que ha sido posible hemos tratado de combinar ambas aproximaciones y comparar los resultados con los obtenidos al aplicarlas independientemente.

Por otro lado, y sobre todo en las soluciones basadas en el aprendizaje automático, siendo el euskera un idioma aglutinante los típicos atributos que se utilizan en este tipo de aproximaciones no son suficientes. Por ello, hemos prestado especial atención a la identificación y selección de atributos que pueden aportar información relevante al respecto. Se ha considerado además que esta identificación y selección de atributos pueda ser válida o al menos ayude a soluciones de otros idiomas de morfología similar.

En general, las tres herramientas de tratamiento de NEs que se han desarrollado dentro de esta tesis han resultado válidas para otras aplicaciones

del área del procesamiento del lenguaje natural, tanto dentro del grupo IXA como fuera de este.

En concreto dentro del grupo IXA, la herramienta de identificación y clasificación de NEs se ha utilizado dentro del marco de los proyectos *HERMES* (Verdejo *et al.*, 2003) y *EPEC* (Aduriz *et al.*, 2006) para el etiquetado automático de diferentes corpus en euskera. La misma herramienta se ha utilizado en las investigaciones realizadas en el ámbito de un sistema de pregunta-repuesta (Alegria *et al.*, 2009).

Fuera del grupo IXA y en el marco del proyecto *OpenTrad*[1] de traducción automática, han sido muy útiles las aportaciones de esta tesis en el trabajo de eliminación de NE en memorias de traducción, sin el cual la publicación de las mismas no hubiera sido posible.

El grupo GALAN de la UPV/EHU, por su parte, ha utilizado también estas aportaciones en el trabajo de investigación realizado acerca de la identificación del dominio de material escolar (Larrañaga *et al.*, 2003).

Por tanto, podemos decir que las aportaciones principales de esta tesis han sido las herramientas fruto de la investigación de las tres liñeas de NEs abordadas. Sin embargo, con este trabajo se han conseguido otras aportaciones que, aunque no sean tan visibles, no son menos importantes. A continuación mencionamos las más relevantes:

- En los experimentos realizados, principalmente en la identificación y clasificación de NEs, hemos confirmado la hipótesis inicial de que combinando métodos simples de diferentes estrategias, como son los métodos basados en el conocimiento de la lengua y los basados en el aprendizaje automático, se pueden desarrollar sistemas robustos, evitando el uso de métodos sofisticados que requieren de muchos más recursos.
- Hemos presentado mecanismos de generación de corpus de evaluación automáticos tanto en la tarea de traducción como en la de desambiguación, explotando características de un corpus periodístico y de la Wikipedia respectivamente.
- Se ha definido una estrategia de enriquecido de diccionarios bilingües estándares para adecuarlos a la tarea de traducción de NE, ya que identificamos que por el hecho de ser parte de una NE, ciertas palabras se comportan de diferente manera y por tanto necesitan un tratamiento

[1] `http://www.opentrad.org/`

especial. Un ejemplo de ello es el elemento *Batuak* de la NE *Nazio Batuak*, que para su traducción en inglés toma la forma de *United*, forma no usual y no contemplada en el diccionario estándar. Con el fin de automatizar ese tratamiento especial surgió la idea de tratarlo a través del enriquecido de diccionarios y así lo hemos hecho, obteniendo mejoras prometedoras. Básicamente el proceso trata de, a partir de una colección de NEs formadas por elementos de este tipo, mapear los elementos que se pueden traducir de cada par de NEs a través del diccionario bilingüe, y tratar de buscar correspondencias entre los elementos que no se han podido traducir. Cuando la búsqueda de correspondencia es exitosa, el par de palabras se incluye en el diccionario, obteniendo así de forma automática un diccionario bilingüe adecuado a la traducción de NEs.

- En la tarea de desambiguación se ha propuesto una modificación en el algoritmo ESA para adecuar el comportamiento de este a recursos pequeños como la Wikipedia en euskera, logrando una mejora importante en los resultados para el euskera.
- Se ha realizado un estudio de la Wikipedia en euskera y de su impacto respecto a Wikipedias más grandes como por ejemplo la del inglés, especialmente en la tarea de desambiguación, obteniendo resultados y estrategias que consideramos aplicables a idiomas con Wikipedias de características similares a las del euskera.
- Una estrategia también portable a otros idiomas es la que se ha propuesto en la tarea de traducción de NEs. Concretamente nos referimos a la estrategia semisupervisada basada en métodos y recurso genéricos como pueden ser las reglas basadas en la distancia de edición y los corpora comparables.

V.2. Conclusiones

Una vez presentadas las principales aportaciones, a continuación se describen las conclusiones más destacables de esta tesis.

Decidir cuándo aplicar técnicas basadas en el conocimiento del idioma o las basadas en el aprendizaje automático no es tarea fácil[2]. El nivel de

[2]Tal y como explicó el profesor Yurafsky en la charla que ofreció en 2011 en la

conocimiento del problema a resolver, los recursos disponibles, el tamaño de los corpora y la idoneidad de los mismos son algunos de los factores más importantes a tener en cuenta para tomar la decisión correcta.

En esta tesis, hemos aplicado ambas aproximaciones para resolver diferentes tareas de NEs, y aunque dependiendo de la tarea, y por supuesto de los recursos, los resultados a veces han variado, en general podemos decir que efectivamente se ha confirmado nuestra hipótesis inicial acerca de que la combinación de sistemas basados en diferentes aproximaciones sea una buena estrategia para mejorar los resultados. Esa mejoría es más notable cuando los sistemas que se combinan son de tipos diferentes, ya que gracias a la complementariedad de los sistemas se consiguen mejoras importantes.

Hemos aplicado tanto técnicas supervisadas como técnicas semisupervisadas y no supervisadas para resolver los problemas planteados en los objetivos. Y gracias a que hemos usado diferentes aproximaciones para resolver el mismo problema, hemos podido analizar y comparar el comportamiento de las mismas, tratando de ver en cada caso cuál era la más idónea. Concretamente, al hacer este análisis en el problema de traducción de las NEs hemos podido ver que, con menos recursos y menor esfuerzo, las estrategias semisupervisadas pueden llegar casi a igualar los resultados de estrategias más costosas como son las supervisadas.

Hemos constatado también que además de necesitar menos recursos y esfuerzos y obtener resultados parejos a los de de sistemas supervisados si para su implementación se utilizan métodos genéricos, los sistemas semisupervisados pueden llegar a ser portables a otros idiomas. Esto es posible gracias a la casi total independencia que se consigue respecto al idioma. Ejemplo de esta portabilidad es el sistema de traducción de las NEs que hemos desarrollado combinando un corpus comparable y la Wikipedia con un sistema de reglas de distancia de edición. Este sistema permite de forma sencilla desarrollar sistemas de traducción para diferentes pares de idiomas. Y así lo hemos demostrado con la evaluación de tres pares de idiomas diferentes que han sido el euskara-castellano, castellano-inglés y euskera-inglés.

En el trabajo realizado entorno a la desambiguación de NEs en euskera, no se han confirmado los buenos resultados de algunas estrategias para idiomas de recursos abundantes, como el caso del inglés, al aplicarlas en entornos de recursos más limitados como el euskera. De hecho, en algunos casos no sólo empeoran los resultados al utilizar Wikipedias de tamaño y características

UPV/EHU. http://www.unibertsitatea.net/blogak/ixa/flirteatzen-ari-da

diferentes, sino que además el comportamiento de algunos métodos cambia completamente. Sistemas que en el caso del inglés se posicionan como los mejores, en el caso del euskera obtienen resultados escasos, como es el caso del algoritmo ESA. Y, al revés, métodos que para el inglés no consiguen resultados destacados en el caso del euskera se sitúan entre los mejores, como por ejemplo el algoritmo UKB. En cualquier caso, y debido al tamaño pequeño de los corpora de evaluación que hemos utilizado, consideramos necesario realizar una evaluación con corpora más grandes y poder así verificar estas conclusiones.

Finalmente el aporte o beneficio de la lematización para resolver problemas de procesamiento del lenguaje natural en euskera no queda claro en este trabajo. No podemos concluir que la pérdida de detalle sobre el contenido y los errores de etiquetado automático creados en el proceso de la lematización compensen siempre la normalización y evidencia estadística que se consigue gracias a ese mismo proceso, ya que el efecto negativo ocasional que se describen Arregi eta Fernández (2002) de lematización en el ámbito de la clasificación automática de textos lo hemos visto repetido en nuestros experimentos de desambiguación. En nuestro caso en concreto, parece que la lematización en lugar de ayudar en el proceso de desambiguación repercute negativamente haciendo que el sistema empeore al aplicarlo sobre datos lematizados.

En cualquier caso, y como anteriormente se ha mencionado, para poder confirmar estas conclusiones consideramos imprescindible una evaluación con corpora más grandes.

V.3. Trabajos Futuros

Las herramientas que resuelven cualquier tipo de problema automáticamente, así es al menos en las que resuelven los problemas del ámbito del procesamiento del lenguaje natural, nunca llegan a la perfección y, por tanto, siempre existen vías de mejora que se pueden abordar. Quizá el abanico de vías de mejora no es tan amplio en el caso de las herramientas basadas en técnicas en el conocimiento, pero sí en el caso de los sistemas basados en técnicas de aprendizaje automático. Estas vías de mejora pueden venir de la mano de un cambio importante como puede ser el uso de algoritmos más complejos para modelar el problema que los utilizados hasta entonces; o de cambios de menor impacto en el diseño del sistema como por ejemplo afinar

la selección de atributos, o ampliar la cantidad de datos para modelar el sistema (Arrieta, 2010), que además de ser más sencillos de llevar a cabo pueden ser también muy útiles en el camino de mejorara de efectividad y calidad de un sistema de este tipo. Por tanto, a raíz de las herramientas generadas en esta tesis podemos decir que queda abierto el camino de mejora de herramientas, principalmente las que resuelven la traducción y desambiguación de las NEs en euskera explotando estrategias semisupervisadas. Consideramos la explotación de la cada vez más usada Wikipedia uno de los primeros pasos a dar para dicha mejoría .

Por otro lado y siguiendo con la Wikipedia consideramos interesante explotar el conocimiento que hemos adquirido de este recurso al utilizarlo en la desambiguación y traducción y trasladarlo a la tarea de identificación y clasificación de NEs, ya que creemos que un uso correcto de características de la Wikipedia puede ayudar a mejorar el sistema NERC, principalmente en la tarea de clasificación.

Al pensar en la mejora de los sistemas es inevitable plantear y prever un análisis y tratamiento de las dudas y conclusiones que han quedado pendientes de resolver en la tarea de la desambiguación. Para ello es necesario repetir los experimentos con mayor cantidad y mejor calidad de datos. Y creemos que en esa mejora de datos va a ayudar a corto plazo la iniciativa de ampliación de la Wikipedia en euskera que se está llevando a cabo. Sin embargo, este aumento de datos no será suficiente para construir un sistema robusto de desambiguación de NEs, ya que en esta tesis no se ha prestado la atención suficiente a resolver los casos NIL (ejemplos sin entrada en la base de conocimiento). Por tanto, además de utilizar la Wikipedia más completa, es necesario diseñar e implementar una estrategia dentro del sistema existente que resuelva adecuadamente esos casos y poder así crear un sistema robusto desde todos los puntos de vista.

Y para finalizar, a pesar de que en esta tesis la identificación y clasificación, la traducción y la desambiguación se han abordado de forma independiente, creemos que la interacción entre ellas pueden resultar de ayuda, y consideramos esa linea de investigación muy interesante para el futuro. Por ejemplo, analizar cómo la traducción y la desambiguación pueden servir para afinar los resultados de un sistema NERC; o cómo los sistemas NERC y NET pueden ayudar en la desambiguación. Es, por tanto, el estudio de esa complementariedad y ayuda entre sistemas es una línea abierta investigación, tomando como punto de partida los sistemas de NEs desarrollados dentro de esta tesis. Y son precisamente los resultados de esa línea de investigación los

que consideramos que tanto desde el punto de vista teórico como práctico pueden traer las mejoras más relevantes en el área.

Bibliografía

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., eta Urizar R. Methodology and steps towards the construction of epec, a corpus of written basque tagged at morphological and syntactic levels for the automatic processing. In Wilson A., Rayson P., eta Archer D., editors, *Corpus Linguistics Around the World.*, 56 lib. of *Book series: Language and Computers*, 1–15, Netherlands, 2006. Rodopi.

Agirre E., Chang A., Jurafsky D., Manning C., Spitkovsky V., eta Yeh E. Stanford-ubc at tac-kbp (knowledge base population). *In Proceedings of the NIST Text Analysis Conference (TAC2009)*, 2009.

Alegria I., Ansa O., Arregi X., Otegi A., eta Soraluze A. Ihardetsi: A question answering system for basque built on reused linguistic processors. *SEPLN-SALTMIL 2009 workshop: Information Retrieval and Information Extraction for Less Resourced Languages.*, 37–43, 2009.

Alegria I., Arregi O., Balza I., Ezeiza N., Fernandez I., eta Urizar R. Design and development of a named entity recognizer for an agglutinative language. *First International Joint Conference on NLP (IJC-NLP-04). Workshop on Named Entity Recognition*, 2004.

Alegria I., Arregi O., Ezeiza N., eta Fernandez I. Lessons from the development of a named entity recognizer. *Procesamiento del Lenguaje Natural*, 25–37, 2006a.

Alegria I., Ezeiza N., eta Fernandez I. Named entities translation based on comparable corpora. *Multi-Word-Expressions in a Multilingual Context Workshop on EACL06.*, 1–8, 2006b.

Alegria I., Ezeiza N., eta Fernandez I. Translating named entities using comparable corpora. *Building and Using Comparable Corpora. LREC 2008 Workshop.*, 2008.

Alegria I., Ezeiza N., Fernandez I., eta Urizar R. Named entity recognition and classification for texts in basque. *II Jornadas de Tratamiento y Recuperación de Información, JOTRI2003*, 2003.

Arregi O. eta Fernández I. Clasificación de documentos escritos en euskara: impacto de la lematización. *I Jornadas de Tratamiento y Recuperación de Información, JOTRI*, 2002.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: kateen eta perpausen identifikazioa eta bere erabilera komazuzentzaile baterako.* Doktoretza-tesia, Lengoaia eta Sistema Informatikoak Saila, Euskal Herriko Unibertsitatea, 2010.

Carreras X., Màrquez L., eta Padró L. A simple named entity extractor using AdaBoost. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, page 155, 2003.

Chinchor N. Overview of MUC-7. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

Fernandez I., Alegria I., eta Ezeiza N. Using wikipedia for named entities translation. *SALTMIL2009 workshop: IR-IE-LRL*, 2009.

Fernandez I., Alegria I., eta Ezeiza N. Semantic relatedness for named entity disambiguation using a small wikipedia. *TSD 2011, LNAI 6836.*, 276–283, 2011.

Larrañaga M., Rueda U., Elorriaga J., eta Arruarte A. Index analysis: a means to acquire the domain module structure. *In proceedings of 10th Conference of the Spanish Association for Artificial Intelligence and 5th Conference on Technology Transfer*, 339–342, 2003.

Verdejo M., Gonzalo J., Márquez L., Padró L., Rodríguez H., eta Agirre E. Hermes, hemerotecas electrónicas: Recuperación multilingüe y extracción semántica. *Jornada de Seguimiento de Proyectos en Tecnologías del Software. Programa Nacional de Tecnologías de la Información y las Comunicaciones*, 2003.