Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 287-290 recibido 21-06-10 revisado 05-07-10 aceptado 06-07-10

Evaluación de la Extracción de Entidades Nombradas de OpenCalais en castellano

Evaluation of Named Entity Recognition in Spanish with OpenCalais

Raquel Toribio
Telefónica I+D
C/ Emilio Vargas 6
28043 Madrid
raquelt@tid.es

Paloma Martínez
César de Pablo-Sánchez
Universidad Carlos III de Madrid
Avda. de la Universidad 30,
28911 Leganés
{pmf,cdepablo}@inf.uc3m.es

Resumen: En los últimos años se han popularizado herramientas de Extracción de Información comerciales dentro del ecosistema de servicios de la Web Semántica. OpenCalais ofrece actualmente reconocimiento y categorización de Entidades Nombradas en castellano de fácil integración en aplicaciones de PLN. Hemos evaluado esta herramienta de anotación de entidades en el corpus de noticias CoNLL 2002. OpenCalais obtiene valores de precisión aceptables en las principales clases (persona, lugares y organización). Sin embargo, en comparación con los prototipos de investigación en castellano puede mejorar la cobertura y el tratamiento de la ambigüedad.
Palabras clave: Reconocimiento y Clasificación de Entidades Nombradas, Evaluación, Extracción de Información, Web Semántica, Castellano

Abstract: The Semantic Web ecosystem has seen the growing popularity of commercial Information Extraction services. Among them, OpenCalais provides Named Entity Recognition and Classification in Spanish. We have evaluated this service in the CONLL 2002 news corpus. The precision results are good enough for the development of applications that use the main classes (person, location and organization). However, recall and the treatment of ambiguous entities could be improved to be in pair with research prototypes.
Keywords: Named Entity Recognition and Classification, Evaluation, Information Extraction, Spanish

1 Introducción[1]

Internet y las tecnologías de la Web 2.0 han propiciado una explosión de la información disponible en diferentes modalidades. Las técnicas de Extracción de Información son una de las alternativas para organizar y mejorar el acceso a este torrente de información. En los últimos años han aparecido varios servicios software comercial que permiten la extracción de palabras claves y de Entidades Nombradas (NE del inglés Named Entity) como OpenCalais[2], Zemanta[3], AlchemyAPI[4], Evri[5], STILUS Sem[6], OpenAmplify[7], SaploTags[8] o BeliefNetworks[9].

Estos servicios se han integrado en numerosas aplicaciones y es previsible que, con el avance del software como servicio, sirvan para mejorar las capacidades semánticas y de interoperabilidad de muchas más en un futuro próximo. Algunas de ellas ya se encuentran disponibles para el procesamiento del castellano como OpenCalais, AlchemyAPI o STILUS Sem.

[1] Este trabajo ha sido parcialmente financiado por la red MA2VICMR (S2009/TIC-1542) y por el proyecto BRAVO (TIN2007-67407-C03-01)
[2] www.opencalais.com
[3] www.zemanta.com
[4] www.alchemyapi.com
[5] www.evri.com
[6] www.daedalus.es/productos/stilus/stilus-sem
[7] www.openamplify.com
[8] saplo.com
[9] www.beliefnetworks.net

El objetivo de este trabajo es analizar el desempeño de la extracción de NE que proporciona OpenCalais para textos en castellano con vistas a su integración en otras aplicaciones de PLN. Para ello hemos utilizado el corpus CoNLL 2002, sobre el que ya se han evaluado varios prototipos de investigación.

2 Descripción de OpenCalais

OpenCalais es un servicio web de Thomson Reuters que permite la extracción de entidades, hechos y eventos de texto libre en inglés, francés y castellano. Su versión en inglés es la que presenta una mayor funcionalidad, si bien en español permite:

- reconocimiento y categorización de entidades usando 15 clases de entidades.
- evaluación de la relevancia de entidades
- desambiguación y enlazado con Linked Open Data para algunos tipos como *Company*.

OpenCalais ofrece un API sencillo que puede ser usado mediante SOAP, REST vía HTTP POST, ó HTTP POST. Como entrada permite documentos de distintos formatos (HTML, HTMLRAW, XML y texto). Además de la etiquetación semántica el servicio incluye la eliminación de cabeceras y otros elementos en HTML así como la detección de idioma.

Como salida ofrece la elección de varios formatos XML/RDF, texto, texto con microformatos o JSON. Los formatos XML/RDF y JSON incluyen URIs derreferenciables que pueden enlazar con una tercera fuente de conocimiento, típicamente Linked Data.

Para la definición de todas las clases utilizadas en OpenCalais existe tanto un esquema RDFS[10] como una ontología OWL[11].

3 Corpus de evaluación CONLL 2002

En nuestra evaluación hemos usado el corpus CONLL-2002 (Sang, 2002) para el castellano. La anotación de NE usa cuatro clases: personas (**PER**), lugares (**LOC**), organizaciones (**ORG**) y otras entidades (**MISC**). Los textos etiquetados proceden de una colección de artículos de la agencia de noticias EFE de Mayo del 2000.

El corpus se desarrolló con el objetivo de evaluar sistemas NERC basados en aprendizaje automático por lo que está compuesto de tres ficheros: *train*, *testa* y *testb*. El primero está destinado al aprendizaje del modelo, *testa* es el conjunto de desarrollo para el ajuste de parámetros y *testb* es el conjunto de test sobre el que se evaluaron los sistemas participantes. Los ficheros contienen 273,037, 54,837 y 53,049 tokens respectivamente.

El formato del corpus es tabular con la tokenización y la segmentación en oraciones pre-establecida. Cada token se corresponde con una línea y las NE están marcadas siguiendo el formato BIO en la columna adyacente. La etiqueta B-XXX se utiliza para identificar el comienzo de una entidad nombrada mientras que el resto de tokens se marcan con I-XXX. XXX se corresponde con la clase de entidad asignada. La etiqueta O se usa para marcar aquellos tokens que no se consideran entidades nombradas. En este esquema de etiquetado se asume que las NEs no son recursivas ni se superponen. En el caso de que una NE esté incluida en otra se marca únicamente la entidad de más alto nivel.

Además del corpus, los organizadores de la tarea proporcionaron un script en Perl destinado a la evaluación. Este script proporciona valores de Precisión, Cobertura (o Recall) y su media armónica F usando $\beta=1$.

4 Diseño de la Evaluación

Para completar la evaluación se ha desarrollado una aplicación que permite salvar las diferencias sintácticas de formato y semánticas entre conjuntos de conceptos y etiquetas.

El conversor CoNLL-texto deshace la tokenización y produce un fichero de texto con una oración por línea. Puesto que el servicio de OpenCalais impone un límite de 100.000 caracteres separan las noticias usando la convención de que una línea con un guión introduce una nueva noticia. Para cada fichero se llama al API SOAP de OpenCalais con un documento de texto plano como entrada y

[10] http://www.opencalais.com/files/RDFS%20schema_09Jun16.txt

[11] http://www.opencalais.com/files/owl.opencalais-4.3a.xml

usando como salida la representación en XML/RDF.

Cada fichero RDF obtenido de OpenCalais contiene información sobre las anotaciones o menciones de NE encontradas en el texto. Por cada mención se proporciona información de offset, longitud, texto y tipo de la mención. OpenCalais permite asignar varias anotaciones a una secuencia de tokens que pueden corresponderse con entidades diferentes. Como esto no es posible en CoNLL, se ha optado por tomar como buena la primera mención que aparezca en el documento RDF. El número de anotaciones con solapamiento o anidamiento han sido 110, un 3,8% de las encontradas por OpenCalais. Aunque este porcentaje es bastante bajo, podría suponer una pequeña mejora sobre los resultados obtenidos si se tomase otro criterio. Como último paso se reproduce la salida en el formato usado en CONLL alineando con la tokenización original.

Como la taxonomía de clases de entidades difiere, para la evaluación se ha utilizado la correspondencia que se muestra en la Tabla 2. Las clases en negrita son de CONLL y en cursiva de OpenCalais. Aun así la correspondencia no es directa pues existen diferencias en los criterios de anotación.

4.1 Evaluación de la detección de menciones

Se ha evaluado el Reconocimiento y Clasificación de Entidades Nombradas según se muestra en la Tabla 1.

Clase	P(%)	R(%)	F(%)	# NE
PER	82,32	67,84	74,38	1007
LOC	62,04	63,05	62,54	1001
ORG	62,59	25,59	36,33	695
MISC	11,19	3,60	5,44	143
ALL	**66.80**	**43.68**	**52.82**	**2846**

Tabla 1: Resultados de la evaluación con CONLL-2002

De la inspección de estos resultados se obtiene que la clase persona (**PER**) es la que presenta los mejores resultados. Para esta clase hay una correspondencia clara y unívoca entre ambas taxonomías. El grado de precisión es aceptable pero sólo es capaz de identificar el 67,84% de las entidades nombradas del fichero de CoNLL.

La precisión para las clases lugar (**LOC**) y organización (**ORG**) es más baja, en el orden del 62%. Sin embargo, mientras que la cobertura para **LOC** se mantiene, para **ORG** es tan solo del 26%. Finalmente, la clase **MISC** presenta unos valores muy bajos pues el recubrimiento mutuo entre ambos esquemas de clasificación es escaso, como se explica con mayor detalle en el siguiente apartado.

Si ignoramos la clasificación y solo evaluamos el reconocimiento, obtenemos los siguientes resultados: P=78,60%, R=51,40%, F=62,16%.

4.2 Evaluación de la detección de tipos

Para realizar un análisis más pormenorizado de las anotaciones, hemos agrupado las diferentes menciones en listas de tipos. Las menciones con el mismo texto y clase de entidad se corresponden con un tipo.

La Tabla 2 muestra el número de tipos anotados por cada clase en el corpus (columna # CONLL) y usando OpenCalais (columna # OpenCalais). Es posible estimar la precisión midiendo el solapamiento entre listas o verdaderos positivos (#TP).

Clase CoNLL	Open Calais	P	R	F
PER	**Person**	**85,1**	**67,8**	**75,5**
LOC	Country	79,2		
	City	63,8		
	ProvOSte	75,0		
	Continent	75,0		
	NatFeat	17,9		
	Region	7,1		
		62,6	**55,9**	**59,0**
ORG	Company	71,7		
	Organizat	59,7		
		62,7	**34,9**	**44,9**
MISC	Currency	0,0		
	MarktIdx	60,0		
	Url	66,7		
		4,1	**1,6**	**2,2**

Tabla 2: Resultados de la evaluación por tipos

El análisis de los resultados y la inspección manual de las listas de tipos indica que el

mapeo de clases como *Region*, *NaturalFeature* o *Currency* no es directo. Las unidades monetarias no aparecen anotadas en CONLL como entidad, del mismo modo que las regiones y otros accidentes naturales que no aparecen explícitamente en mayúsculas (*oeste kosovar*).

El reconocimiento de personas presenta problemas con los nombres formados por iniciales y los nombres incompletos. Por ejemplo, en el mismo documento se reconoce *Sofia Loren* pero no *Sofia*. Otros problemas recurrentes son la segmentación de entidades adyacentes, como en la frase "*ex dirigente de Euskadiko Ezkerra Kepa Aulestia",* las enumeraciones o los títulos en mayúsculas.

El descenso en la precisión de **LOC** se explica porque es más genérico que la unión de todas las clases de OpenCalais englobadas. OpenCalais no detecta entidades como direcciones, calles, avenidas, establecimientos, etc. Por otro lado, de la inspección visual de los diccionarios extraídos para los tipos *City* y *ProvinceOrState* se observa que existen bastantes errores de clasificación entre ambos. Por ejemplo *Castilla La Mancha, Cataluña y* otros nombres de autonomías son asignados a la clase *City* mientras que otras entidades ambiguas como *Alicante*, *Soria* o *León* lo hacen como *ProvinceOrState* independientemente de su contexto.

De los dos tipos de entidades de OpenCalais clasificados como **ORG**, la clase *Company* obtiene mejor precisión, posiblemente porque se usa una base de datos propietaria de Thomson que contiene un listado de empresas registradas. Tiene dificultades para encontrar organizaciones definidas por siglas cuando no han sido identificadas con su nombre completo anteriormente, nombres de equipos deportivos, y organizaciones que pueden corresponderse también con lugares como países y ciudades.

La baja cobertura de la clase **MISC** se atribuye a que engloba entidades, tales como nombres de películas y leyes, que no se corresponden con ninguna de las clases de OpenCalais. Otros, como asambleas y ferias, los clasifica dentro de la categoría de *Organization*. Por último, la inspección manual de *Url*, *MarketIndex*, *Region* y *Continent* ha proporcionado resultados de 100% para la precisión.

5 Conclusión

La Tabla 3 contrasta los resultados obtenidos para el servicio de anotación de entidades de OpenCalais en el corpus CONLL-02 con algunos de los mejores resultados de prototipos de investigación. Aunque la comparación no es estrictamente justa, ya que los prototipos han sido optimizados para esta tarea, refleja que existe un margen para la mejora especialmente en la cobertura del sistema.

Sistema	P (%)	R (%)	F
Ferrández, 2006	83.34	83.41	83.37
Carreras, 2002	81.38	81.40	81.39
Florian, 2002	78,70	79,40	79,05
OpenCalais	**66.80**	**43.68**	**52.82**
baseline	26.27	56.48	35.86

Tabla 3: Resultados de la evaluación con CONLL-2002

Aun así, OpenCalais ofrece una alternativa robusta y fácil de integrar en otras aplicaciones junto a la posibilidad de enlazar las entidades reconocidas con descripciones estructuradas de la entidad. OpenCalais ofrece un mayor número de funcionalidades y posiblemente mejores resultados para el inglés como esperamos comprobar en trabajos futuros.

Bibliografía

Carreras, X., L. Márquez, y L. Padró, 2002. Named Entity Extraction using AdaBoost. Proceedings de CoNLL-2002 (pp. 167-170), Taipei, Taiwan.

Ferrández, O, A. Toral, y R. Muñoz. 2006. Fine Tuning Features and Post-processing Rules to Improve Named Entity Recognition. NLDB 2006 (pp. 176–185)

Florian R. 2002. Named Entity Recognition as a House of Cards: Classifier Stacking. *Proceedings of CoNLL-2002 (pp. 175-178)*, Taipei, Taiwan, 2002.

Sang, E. T. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002* (pp. 155-158). Taipei, Taiwan