Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 13-20 recibido 30-10-2016 revisado 13-01-2017 aceptado 12-02-2017

# POL: un nuevo sistema para la detección y clasificación de nombres propios*

## *POL: a new system for named-entity detection and categorisation*

**Rogelio Nazar, Patricio Arriagada**
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
Avenida El Bosque 1290, Viña del Mar, Chile
rogelio.nazar@pucv.cl, patricio.arriagada.s@mail.pucv.cl

**Resumen:** El objetivo de este trabajo es desarrollar una metodología para la detección y clasificación de nombres propios (NP) en las categorías de antropónimo, topónimo y nombre de organización. La hipótesis sobre la que se basa la investigación es que el contexto de aparición de los NP –definido como las $n$ palabras previas– así como los elementos que componen el NP mismo, pueden aportar pistas para predecir el tipo de entidad. Para tal fin, se diseñó un algoritmo de clasificación supervisado que se entrena con un corpus ya anotado por otro sistema, que en el caso de nuestros experimentos fue la suite de analizadores de idiomas FreeLing anotando el corpus de la Wikipedia en castellano. En el entrenamiento, nuestro sistema aprende a relacionar tipos de entidades con palabras del contexto así como las que componen los NP anotados. Se evalúan los resultados en el corpus CONLL-2002 y también con un corpus de geopolítica perteneciente a la revista *Le Monde Diplomatique* en su edición en castellano. Se compara además el desempeño en ese corpus de distintos sistemas de extracción y clasificación de NP en castellano.
**Palabras clave:** Entidades nombradas, nombres propios, lingüística textual

**Abstract:** The purpose of this research is to develop a methodology for the detection and categorisation of named entities or proper names (PPNN), in the categories of geographical place, person and organisation. The hypothesis is that the context of occurrence of the entity –a context window of $n$ words before the target– as well as the components of the PN itself may provide good estimators of the type of PN. To that end, we developed a supervised categorisation algorithm, with a training phase in which the system receives a corpus already annotated by another NERC system. In the case of these experiments, such system was the open-source suite of language analysers FreeLing, annotating the corpus of the Spanish Wikipedia. During this training phase, the system learns to associate the category of entity with words of the context as well as those from the PN itself. We evaluate results with the CONLL-2002 and also with a corpus of geopolitics from the journal *Le Monde Diplomatique* in its Spanish edition, and compare the results with some well-known NERC systems for Spanish.
**Keywords:** Named entities, proper names, text linguistics

## *1 Introducción*

El presente artículo sintetiza una investigación en curso cuyo objetivo es desarrollar una metodología de detección y clasificación de nombres propios (NP) o *named entities*, tal como se conocen generalmente en la comunidad del Procesamiento del Lenguaje Natural (PLN). La tarea se descompone en dos fases que se resuelven de manera independiente pero consecutiva: en primer lugar la detección del NP, que incluye el problema de su delimitación, y luego su clasificación en las categorías de antropónimos, topónimos y nombres de organizaciones. Identificamos como

* Investigación financiada por CONICYT (Gobierno de Chile), Proyecto Fondecyt 11140686: "Inducción automática de taxonomías de sustantivos generales y especializados a partir de corpus textuales desde el enfoque de la lingüística cuantitativa". Agradecemos también a los revisores por sus comentarios.

ISSN 1135-5948 

POL a nuestro algoritmo por Persona, Organización, Lugar.

Basándonos en el trabajo previo de Arriagada (2016), la hipótesis en la que la investigación se basa es que el contexto de aparición de los NP –definido como las $n$ palabras previas– además de las palabras que componen el NP mismo, pueden aportar elementos predictores del tipo de entidad. Por ejemplo, un verbo, un adverbio o un adjetivo, pueden ofrecer pistas respecto al tipo de entidad que acompañan cuando están en una relación predicado/argumento con un NP (De Miguel, 1999). Así, cabe esperar que el sujeto de un verbo como *considerar* sea un sujeto humano. Incluso sin análisis sintáctico, se hipotetiza que la mera coocurrencia de estos elementos puede aportar información útil.

El diseño experimental para poner a prueba esta hipótesis se basa en el desarrollo de un algoritmo supervisado, con una etapa de entrenamiento en que recibe texto anotado por otro sistema, y una etapa de prueba en la que detecta y clasifica entidades en texto sin anotar. Durante el entrenamiento, el algoritmo aprende a relacionar palabras que componen los NP y las que aparecen en el contexto de estos con la categoría de entidad asignada por el otro sistema. En el caso de los experimentos descritos en este artículo, ese sistema es la suite de analizadores de idiomas FreeLing[1] (Carreras et al., 2004), pero podría utilizarse otro sistema sin inconveniente y, al menos en teoría, incluso una anotación manual, aunque esto sería poco práctico debido a los volúmenes de datos que se requieren. En los experimentos aquí descritos, se utilizó como corpus de entrenamiento la Wikipedia en castellano.

Nuestro enfoque también asume el supuesto de partida de la lingüística textual de la correspondencia entre un referente y un texto, es decir que todas las menciones de un NP en un texto refieren a una misma entidad. Por tal motivo, el análisis procede con un texto a la vez y no un corpus como aglomeración de varios textos.

Los resultados muestran una tasa de acierto en las tareas de detección y clasificación de entidades comparables a las de los sistemas más conocidos. Cabe destacar, además, la simplicidad de la metodología ya que, al contrario de herramientas como FreeLing, no utiliza ningún tipo de pre ni post procesamiento de los textos. No requiere lematización, etiquetado morfológico ni análisis sintáctico.

En el sitio web que acompaña a este artículo se ofrece una implementación del algoritmo en código abierto, en forma de dos scripts Perl, uno para entrenamiento y otro para clasificación. Los dos scripts son muy sencillos, con menos de 100 líneas cada uno. Además del código, ofrecemos una demo en línea en la que se puede comparar el desempeño con otros clasificadores: http://www.tecling.com/pol

## *2 Marco teórico*

El NP ha sido estudiado por disciplinas tan diversas como la geografía lingüística, la filosofía, la gramática comparada, la lexicología y la traductología, entre otras. Ofrecemos a continuación una breve caracterización del NP y una síntesis del trabajo relacionado que se ha producido en el ámbito del PLN.

### 2.1 Características de los NP

En la teoría lingüística existe consenso respecto a algunos rasgos prototípicos que caracterizan al NP (Fernández Leborans, 1999; RAE, 2009), entre los que podemos encontrar los siguientes: 1) **Introducción mediante mayúscula**: al menos en castellano y otras lenguas europeas, el NP se distingue gráficamente del nombre común (NC) por medio de este rasgo[2]; 2) **Flexión fija**: el NP se distingue en general del NC por carecer de flexión, salvo en casos muy específicos como los comentados por Coseriu (1982) donde el plural de los NP es indicio de un uso que se asimila al de un NC (ej.: *En la galería se exhiben varios Picassos*); 3) **Monorreferencialidad o unicidad referencial**: sería un defecto de forma que en un mismo texto un NP hiciera referencia a entidades distintas sin advertencia por parte del autor. Esto conlleva el hecho de que el NP no requiera el uso de artículo (definido o indefinido) y que de hecho su uso sin artículo sea normativo en todos los contextos (**El Jorge Luis Borges...*[3]); 4) **Falta de significado léxico**: los NP no poseen significado sino referencia (Russell, 1905) y

[1] http://nlp.lsi.upc.edu/freeling/

[2] Esto no es así en alemán, donde la mayúscula inicial es rasgo distintivo de la categoría gramatical nombre en general (*Das Auto*).

[3] A pesar de esto, el habla coloquial en Barcelona o regiones como Cuyo en Argentina registra el uso de artículo en el NP, posiblemente por influencia del catalán.

por esta razón no aparecen en los diccionarios lexicográficos, a menos que su uso pase a denominar una clase en lugar de un particular, como suele suceder por ejemplo en el caso de las marcas comerciales (*Comprar un kleenex*); 5) **Imposibilidad de traducción exacta**: la traducción o transliteración de los NP es a menudo problemática a menos que se trate de nombres ya convencionalizados (ej. *Miguel/ Michael/ Mijail*).

Estos o algunos de estos rasgos, que son inherentes a los NP, podrían ser objeto de interés para detectarlos y clasificarlos. Otros rasgos, aunque menos estudiados, podrían servir también, como por ejemplo un patrón de coocurrencia característico del NP. Esta es una de las variables que se pretende explorar en este trabajo, como indicador del valor o función referencial de una unidad léxica.

## 2.2 Detección/clasificación de NP

En el campo del PLN, el reconocimiento y la clasificación de NP tiene una larga historia en aplicaciones de recuperación o extracción de información y en la búsqueda de respuestas, dentro de la especialidad del *named entity recognition & classification* (Manning, Raghavan, y Schütze, 2008). Existen en la actualidad diversos sistemas para detectar y clasificar los NP, que en esta comunidad disciplinaria se denominan de manera genérica "entidades nombradas". Las categorías son principalmente las de antropónimo, topónimo o nombre de organización, aunque en rigor las entidades nombradas incluyen también fechas y signos diversos para designar valores o cantidades. Sin embargo, las clasificaciones pueden ser incluso más finas, como las exhibidas por Nadeau y Sekine (2007).

Un antecedente importante en la historia de este tipo de sistemas es el trabajo realizado en las Message Understanding Conferences (MUC) celebradas desde 1987, donde se demuestra que el reconocimiento de entidades es un componente fundamental de los sistemas de extracción de información, que requiere tanto del análisis léxico, sintáctico y en algunos casos hasta textual (Grishman y Sundheim, 1996; Wilks, 1998). Típicas estrategias utilizadas entonces y en la actualidad son el uso de *trigger-words* o palabras asociadas a un tipo de entidad, como "Inc." en el caso de las corporaciones, y las *gazetteer*, o listados de NP de distinto tipo.

Otro hito particularmente importante en el caso de la lengua castellana fue la celebración de la *Conference on Computational Natural Language Learning* (CoNLL-2002) por la tarea de reconocimiento de entidades utilizando un corpus de la agencia de noticias EFE (Tjong Kim Sang, 2002). Este corpus se convirtió en un estándar para medir el desempeño de los clasificadores que aparecieron después, e impuso la notación BIO (Ramshaw y Marcus, 1995; Tjong Kim Sang y Veenstra, 1999) para determinar el comienzo de un NP (B), su interior (I) o su finalización (0).

Carreras, Màrquez, y Padró (2002) obtuvieron el mejor desempeño en esta competencia con un algoritmo de aprendizaje automático (*AdaBoost*) y utilizando distintas pistas tales como las palabras del contexto, las categorías gramaticales, rasgos ortográficos, así como las *trigger-words* y *gazetteers*. Reportaron un 79 % de precisión y cobertura en la detección y un 81 % también en ambos valores en la clasificación, valores cercanos a los de algunos competidores como Florian (2002) con 78 % de precisión y 79 % de cobertura en la tarea de clasificación. Padró y Padró (2005) intentaron un modelo más económico en términos computacionales, basado en el algoritmo *Causal-States Splitting Reconstruction*, aunque sin llegar a superar esos resultados. Gamallo et al. (2014), presentando el programa CitiusNEC, describen una propuesta más compleja y de alto coste computacional, que integra diversas estrategias y que, sin embargo, tampoco superan los resultados informados por Carreras, Màrquez, y Padró (2002). Agerri, Bermudez, y Rigau (2014) son los que más se acercan, con una propuesta basada en el algoritmo de aprendizaje automático *Maximum Entropy*. Señalamos, sin embargo, el riesgo de evaluar siempre con ese mismo corpus ya que cuando se utilizan corpus distintos, y especialmente de otros géneros, las cifras de desempeño varían significativamente (van Hooland et al., 2015).

Se ha avanzado mucho desde las primeras conferencias MUC, y sin embargo el nicho de investigación continúa abierto hasta la actualidad, pues los sistemas de análisis aún no alcanzan su plenitud de desarrollo. Quedan además algunas lagunas. Si bien disponemos de algunos análisis detallados como el de Tkachenko y Simanovsky (2012) en corpus en inglés, queda por hacer un estudio del

peso relativo que tienen los distintos elementos gramaticales que acompañan al NP, como sería el caso de la colocación verbo-nominal, línea que sigue abierta a pesar de que Grishman y Sundheim (1996) ya la mencionaban. Otro vacío que se puede percibir en la literatura es la escasez de estrategias basadas en el enfoque de la lingüística textual, ya que en general no consideran al texto como unidad comunicativa con sentido completo (van Dijk, 1992). En el sentido más elemental, de esto se desprende por ejemplo que si una entidad es mencionada más de una vez dentro de un texto, cabe esperar que la referencia sea la misma en cada mención.

De todo el trabajo relacionado, el de Solorio (2004) es el que más se parece al presente artículo. Ella también utiliza FreeLing como una "caja negra" con la cual entrenar un clasificador. En su caso, el clasificador es *Support Vector Machines* y las pistas para el entrenamiento son las categorías gramaticales de las dos palabras anteriores y posteriores a la palabra analizada. Se puede decir, por tanto, que en este artículo exploramos y desarrollamos algunas consecuencias y variantes de esta idea.

La Tabla 1 exhibe los sistemas que utilizamos para evaluar el desempeño de nuestro método. Además de FreeLing y CitiusNEC, son Stanford-NERC (Finkel, Grenager, y Manning, 2005), una implementación de un algoritmo basado en el muestreo de Gibbs, y Semantria, un producto comercial de la empresa Lexalytics[4], basado en aprendizaje automático (*Conditional Random Fields*), expresiones regulares y extensas bases de datos de NP de distinto tipo.

| Sistema | URL |
|---|---|
| CitiusNEC | gramatica.usc.es/pln |
| FreeLing 3.1 | nlp.lsi.upc.edu/freeling |
| FreeLing 4.0 | nlp.lsi.upc.edu/freeling |
| Semantria | semantria.com |
| Stanford | nlp.stanford.edu/software |

Tabla 1: sistemas utilizados en la evaluación

## *3 Metodología*

La metodología se basa en el diseño de un algoritmo de clasificación supervisado que, en una etapa de entrenamiento, recibe un corpus anotado y, en la etapa de prueba, utiliza este aprendizaje para detectar y clasificar NP en corpus sin anotar. Describimos a continuación materiales y procedimiento.

### 3.1 Materiales utilizados

Los materiales utilizados son un etiquetador de nombres propios, en nuestro caso FreeLing, y un corpus para utilizar como entrenamiento, en este caso la Wikipedia en castellano[5] (aprox. 586 millones de tokens). Cabe aclarar que esto incluye solamente el texto, es decir que se descartan todos los metadatos y la estructura de categorías de este recurso. No es necesario, por tanto, que el algoritmo se entrene únicamente con la Wikipedia, ya que podría ser entrenado con cualquier otro corpus que tenga un tamaño similar. Además de este recurso, también realizamos pruebas con el corpus de enegramas de Google Books (Lin et al., 2012) como una fuente para extraer nombres propios, aunque esto no es indispensable. Recalcamos que estos recursos solo se requieren durante la fase de entrenamiento ya que, una vez entrenado, el sistema funciona de manera completamente independiente.

Para la evaluación utilizamos el ya mencionado corpus CONLL-2002 (53.049 tokens) y un segundo corpus que confeccionamos a partir de una muestra intencionada de 8 artículos del periódico *Le Monde Diplomatique* en su edición en castellano (21.205 tokens). Las razones para utilizar este segundo corpus son diversas. En primer lugar, el corpus CONLL-2002 es menos apropiado para nuestra evaluación porque nuestro sistema está diseñado para analizar textos individuales. Otro motivo es disponer de un corpus de tamaño más reducido que nos permitiera tener mayor control para un análisis cualitativo y un control riguroso de lo que ocurre, ya que el corpus de CONLL-2002 contiene una considerable tasa de error en el etiquetado. Finalmente, un corpus de naturaleza tan distinta al de entrenamiento muestra un escenario más realista. En este caso, un corpus de geopolítica es especialmente exigente por las múltiples referencias a figuras políticas, territorios y organizaciones en situaciones diversas.

### 3.2 Procedimiento

La metodología se divide primero en fases de entrenamiento y clasificación, y luego cada

[4]https://www.lexalytics.com/

[5]https://dumps.wikimedia.org/eswiki/20160920/

una de estas se divide a su vez en dos.

### 3.2.1 Fase de entrenamiento

El entrenamiento tiene una primera fase más simple de recolección de NP y una segunda en la que asocia la categoría asignada por el otro sistema a cada NP con sus componentes y con elementos del contexto.

**a) Extracción de un listado de NP**: el primer paso consiste en obtener un listado de nombres propios, para lo cual utilizamos el corpus de enegramas Google Books (Lin et al., 2012) debido a su gran tamaño y libre disponibilidad. Para extraer los nombres propios se definió un coeficiente $r$ (ecuación 1) que expresa la razón entre la frecuencia de una unidad léxica $j$ escrita con mayúscula inicial ($Mj$) y la frecuencia total de esa misma unidad ($Tj$).

$$r_j = \frac{f(M_j)}{f(T_j)} \tag{1}$$

$$r_j \geq u \rightarrow j \in NP \tag{2}$$

Luego, (2), se consideró como perteneciente al conjunto de los NP a cualquier unidad $j$ con un coeficiente $r$ superior a un umbral arbitrario $u$[6].

**b) Registro de las palabras del contexto**: el paso anterior permite extraer listados de nombres propios, pero nuestro método requiere también asociar estos nombres a una categoría. Como ya se mencionó, esta parte del aprendizaje consiste en relacionar el vocabulario que aparece en el contexto del NP (en una ventana de $n$ palabras) con los tipos de entidades. Para esta parte del entrenamiento utilizamos el corpus de la Wikipedia etiquetado previamente con el clasificador de entidades de FreeLing. A modo ilustrativo, considérese la Tabla 2 con un ejemplo de un fragmento de texto etiquetado.

Una considerable proporción de las apariciones de *Ottawa* y *Toronto* contendrán la palabra *ciudad* en su contexto inmediato, y lo mismo ocurrirá en el caso de muchas otras ciudades. Esto sugiere que la aparición de esta palabra en el contexto de estos NP servirá como predictor de que dicho NP sea un topónimo.

| Forma | Lema | Etiqueta |
|---|---|---|
| Su | su | DP3CS0 |
| capital | capital | NCFS000 |
| es | ser | VSIP3S0 |
| la | el | DA0FS0 |
| ciudad | ciudad | NCFS000 |
| de | de | SPS00 |
| Ottawa | ottawa | NP00G00 |
| y | y | CC |
| la | el | DA0FS0 |
| ciudad | ciudad | NCFS000 |
| más | más | RG |
| poblada | poblar | VMP00SF |
| es | ser | VSIP3S0 |
| Toronto | toronto | NP00G00 |
| . | . | Fp |

Tabla 2: ejemplo del análisis de FreeLing en la clasificación de entidades

Del entrenamiento se deriva entonces un modelo que registra la frecuencia de aparición del vocabulario en el entorno de cada clase de NP. Podemos representar este modelo como una estructura de datos dividida en las tres claves que hemos considerado: $P$ para la categoría de persona, $O$ para organización y $L$ para lugar (opcionalmente se puede incluir la categoría "otros"). Así, cuando en el entrenamiento el sistema encuentra una instancia de NP a la que se ha asignado una determinada categoría, se registrarán en cada caso las palabras que ocurren en las $n$ posiciones anteriores, si es que estas no pertenecen a la lista de exclusión $S$, que contiene signos de puntuación y gramemas de alta frecuencia.

La función $L$ (3) registra los elementos contextuales indicadores de lugar, aunque la representación es la misma en las tres categorías.

$$L\,(j) = \sum_{i=1}^{n} \begin{cases} 1 \text{ if } j \in C_i \wedge j \notin S \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Por cada elemento $i$ del conjunto $C$ de contextos de aparición de los NP de categoría $L$ en el entrenamiento, con $n = \text{card}(C)$, se registra la frecuencia del elemento contextual $j$, que podría ser por ejemplo la palabra *ciudad*. Por simplicidad, representamos el contexto $C_i$ como un conjunto y no como se-

[6] En la versión original del artículo definíamos un umbral $u$ de 0.8, lo que resulta en un algoritmo muy conservador (Tabla 3). Por sugerencia de los evaluadores, experimentamos con $u = 0$ y obtuvimos un mejor equilibrio entre precisión y cobertura. No eliminamos esta parte del artículo, sin embargo, porque podría darse un escenario donde interese más precisión que cobertura.

cuencia de palabras, ignorando por tanto posición y distancia. Registramos además solo la frecuencia absoluta. No hay normalización, lematización ni etiquetado morfosintáctico. Solo se registran las formas flexionadas de las palabras tal como aparecen en el corpus sin distinguir categorías gramaticales.

**c) Registro de componentes del NP**: Además de las palabras del contexto, de manera independiente el algoritmo asocia tipo de entidad con los componentes del mismo NP, con la misma mecánica que en el caso de los elementos del contexto. En el caso de un antropónimo, por ejemplo, tendríamos la distinción entre nombres de pila y apellidos. De esta manera, si en el corpus de entrenamiento se ha observado en reiteradas ocasiones que el componente *Vincent* forma parte de entidades que han sido clasificadas como antropónimos, entonces el sistema tendrá indicios para clasificar luego un nombre como *Vincent Heredia*, aunque este último nunca haya aparecido en el corpus de entrenamiento.

### 3.2.2 Fase de detección y clasificación

A partir de texto no anotado, la primera fase consiste en la detección de los NP, delimitando principio y final, para su posterior clasificación en las categorías ya mencionadas.

Para la primera parte se analiza de manera secuencial cada palabra del corpus de entrada, definida como una pieza léxica entre espacios en blanco o signos de puntuación. Si una palabra aparece con mayúscula inicial y no cumple un patrón de número romano ni adverbio y no ha recibido normalmente etiquetas distintas a la de NP en el corpus de entrenamiento, se declara el comienzo de un posible NP. Si la siguiente palabra comienza con minúscula y no es un gramema, entonces se delimita el final del candidato a NP. Esto permite la detección correcta de secuencias como *Ciudad del Cabo*, *Isla de Pascua*, etc. De estar definido el parámetro $u$, serían descartados los candidatos que no tengan al menos un componente que haya sido visto formando parte de un NP en el corpus de entrenamiento.

Luego, la etapa de clasificación consiste en dos partes. Primero, por cada componente $j$ del NP, se seleccionará la categoría $K_j$ que registre la frecuencia más alta para ese elemento (4). Así, si $j$ es $Vincent$, entonces $P_j$ registrará probablemente el valor más alto.

$$K_j = \max(P_j, O_j, L_j) \tag{4}$$

La segunda etapa es idéntica a la anterior, con la única diferencia de que no son las palabras del NP mismo las que se consideran sino las que aparecen en las $n$ posiciones anteriores. Es en este punto donde cobra importancia tomar el texto completo como unidad de análisis en lugar de usar un contexto únicamente oracional, porque esto quiere decir que se estudian todas las menciones de la entidad en el texto y se dispone de más elementos para tomar la decisión sobre su categoría.

Para la decisión final se procede por votación simple, lo cual trae aparejeado el problema de los empates. En el caso de un NP de dos palabras, cada componente podría ser clasificado de manera distinta. Y siendo dos clasificadores (contexto y componentes) hay casos en que un mismo NP es clasificado de manera distinta. Una función de desempate calcula la certeza $k$ de cada clasificador como la razón entre su primera y su segunda opción ($k = \frac{O_1}{O_2}$), y otorga precedencia el clasificador que se muestre más seguro.

## *4 Resultados*

Presentamos a continuación una evaluación comparativa de los resultados tomando como referencia los sistemas presentados en otros trabajos. Evaluamos primero midiendo porcentajes de precisión y cobertura en la detección de NP en el corpus CONLL-2002 en comparación con los de sistemas más conocidos, tal como son reportados por Gamallo et al. (2014). La Tabla 3 informa los valores totales que se obtienen con el script *conlleval.pl* que proveen los organizadores de CONLL-2002, y el distinto desempeño según cómo se ajusten los parámetros $u$ y $n$.

| Sistema | precisión | cobertura | F1 |
|---|---|---|---|
| CitiusNEC | 67.47 | 66.33 | 66.89 |
| FreeLing 3.1 | 75.08 | 76.60 | 75.98 |
| POL $u = 0{,}8$ | **90.24** | 53.56 | 67.21 |
| POL $\neg u \wedge n = 1$ | 76.04 | 81.52 | 78.68 |
| POL $\neg u \wedge n = 2$ | 75.95 | 82.78 | 79.22 |
| POL $\neg u \wedge n = 3$ | 76.06 | 83.28 | **79.51** |
| POL $\neg u \wedge n = 4$ | 76.03 | **83.33** | 79.51 |
| OpenNLP | 78.96 | 79.09 | 79.02 |

Tabla 3: Evaluación comparativa de los diversos sistemas para la detección de NP en el corpus CONLL-2002

En general se aprecia que los resultados son similares a los informados por otros autores, excepto cuando se define un $u$ alto, que resulta en alta precisión y baja cobertura.

Repetimos los experimentos utilizando el segundo corpus de evaluación, el de la revista *Le Monde Diplomatique*, y comparamos el desempeño en ese corpus con los cinco sistemas NERC mencionados en la Tabla 1. Para ello, los autores detectamos y clasificamos manualmente todos los NP en este corpus, revisando mutuamente nuestra anotación para evitar posibles casos de desacuerdo. Este procedimiento arrojó un total de 537 entidades distintas. La Tabla 4 muestra los resultados de la detección (Pol con $n = 3$). Se aprecia que en general los valores son comparables, con un patrón similar al caso anterior. POL con $u = 0{,}8$ es preciso pero excesivamente conservador, y Semantria muestra esta tendencia de manera incluso más acusada.

| **Sistema** | **precisón** | **cobertura** | **F1** |
|---|---|---|---|
| CitiusNEC | 67 | 88 | 76 |
| FreeLing 3.1 | 74 | 88 | 80 |
| FreeLing 4.0 | 77 | **88** | **82** |
| POL $u = 0{,}8$ | 83 | 66 | 73 |
| POL $\neg u$ | 74 | 79 | 76 |
| Semantria | **94** | 38 | 54 |
| Stanford | 73 | 76 | 74 |

Tabla 4: Evaluación comparativa de los diversos sistemas para la detección de NP en el corpus de *Le Monde Diplomatique* en castellano

| **Sistema** | **Correctos** | **Totales** | **Precisión** |
|---|---|---|---|
| CitiusNEC | 336 | 477 | 70 |
| FreeLing 3.1 | 358 | 476 | 75 |
| FreeLing 4.0 | 359 | 477 | 75 |
| POL $u = 0{,}8$ | 273 | 358 | 76 |
| POL $\neg u$ | 302 | 429 | 70 |
| Semantria | 179 | 205 | **87** |
| Stanford | 313 | 409 | 76 |

Tabla 5: Evaluación comparativa de los diversos sistemas en la clasificación de las entidades antes detectadas en el corpus de *Le Monde Diplomatique* en castellano

En la Tabla 5 se exhiben los resultados de la tarea de clasificación de las entidades que han sido previamente detectadas por cada sistema. En este caso también, Semantria lleva la delantera en precisión, aunque la cobertura es muy escasa. El resto de los sistemas muestra una precisión similar.

Las limitaciones de espacio impiden la inclusión de más variantes en los parámetros, pero las que probamos, como el uso del contexto derecho, no ofrecen mejores resultados. Tampoco podemos ofrecer un análisis pormenorizado de los errores que pudimos observar en los resultados de los distintos sistemas evaluados, aunque en general podemos resaltar, en la tarea de detección, la dificultad de desambiguar palabras que se introducen con una mayúscula pero no son NP ni ejercen función de sustantivo en la oración, lo que sucede particularmente en CitiusNEC y las dos versiones de FreeLing, estas dos últimas muy similares entre sí.

En cuanto a la tarea de clasificación, son fuente de problemas los casos de polisemia regular, donde resulta difícil distinguir por ejemplo si un NP actúa como un topónimo o nombre de organización. En al caso de nuestro algoritmo, el error más frecuente es clasificar nombres de organizaciones que también aplican a nombres de personas, como *Louis Vuitton*.

## *5 Conclusiones y trabajo futuro*

Hemos presentado una metodología para detección y clasificación de NP por medio de un algoritmo que se entrena con los resultados de otro sistema. La propuesta se caracteriza por su simplicidad y bajo coste computacional al no requerir ningún procesamiento del corpus. El corpus CONLL-2002 se procesa en menos de un segundo en un PC de escritorio, lo cual representa una ventaja importante. En cuanto a desempeño, la evaluación comparativa demuestra que es comparable al de otros sistemas más complejos. Los resultados sugieren que un enfoque cuantitativo de este tipo con entrenamiento y test es adecuado, ya que extrae los datos directamente del corpus y no de la introspección del investigador, como los modelos basados en reglas. La simplicidad del método, además, facilita su portabilidad a otras lenguas con menos recursos.

Como trabajo futuro, debemos continuar experimentando con distintos tamaños y tipos de ventanas de contexto y de corpus de entrenamiento, así como reproducir los experimentos en otras lenguas. Otra vía de investigación es estudiar el poder predictor de las unidades del contexto en función de su categoría gramatical y también en función de las relaciones sintácticas que contraen los elementos del contexto con el NP.

## *Bibliografía*


Agerri, R., J. Bermudez, y G. Rigau. 2014. Ixa pipeline: Efficient and ready to use multilingual nlp tools. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, páginas 3823–3828, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arriagada, P. 2016. *Análisis y clasificación de nombres propios en artículos de geopolítica de la revista Le Monde Diplomatique: una aproximación desde la gramática del texto*. Tesis de grado, Pontificia Universidad Católica de Valparaíso.

Carreras, X., I. Chao, L. Padró, y M. Padró. 2004. Freeling: An open-source suite of language analyzers. En *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, páginas 239–242.

Carreras, X., L. Màrquez, y L. Padró. 2002. Named entity extraction using adaboost. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 167–170, Stroudsburg, PA, USA. ACL.

Coseriu, E. 1982. El plural en los nombres propios. En *Teoría del Lenguaje y Lingüística General*. Gredos, Madrid, páginas 261–281.

De Miguel, E. 1999. El aspecto léxico. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española*. Espasa Calpe, Madrid, páginas 2977–3060.

Fernández Leborans, M. J. 1999. El nombre propio. En I. Bosque y V. Demonte, editores, *Gramática descriptiva de la lengua española*. Espasa Calpe, Madrid, páginas 77–128.

Finkel, J. R., T. Grenager, y C. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. En *Proceedings of the 43nd Annual Meeting of the ACL*, páginas 363–370.

Florian, R. 2002. Named entity recognition as a house of cards: Classifier stacking. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 1–4, Stroudsburg, PA, USA. ACL.

Gamallo, P., J. C. Pichel, M. Garcia, J. M. Abuín, y T. Fernández-Pena. 2014. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno big data. *Procesamiento del Lenguaje Natural*, (53):17–24.

Grishman, R. y B. Sundheim. 1996. Message understanding conference-6: a brief history. En *16th International Conference on Computational Linguistics*, páginas 466–471.

Lin, Y., J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, y S. Petrov. 2012. Syntactic annotations for the google books ngram corpus. En *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, páginas 169–174, Stroudsburg, PA, USA. ACL.

Manning, C. D., P. Raghavan, y H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Nadeau, D. y S. Sekine. 2007. A survey of named entity recognition and classification. *Journal of Linguisticae Investigationes*, 30(1):1–20.

Padró, M. y L. Padró. 2005. A named entity recognition system based on a finite automata acquisition algorithm. *Procesamiento del Lenguaje Natural*, (35):319–326.

RAE. 2009. *Nueva gramática de la lengua española*. Espasa Libros, Madrid.

Ramshaw, L. A. y M. P. Marcus. 1995. Text chunking using transformation-based learning. En *Third Workshop on Very Large Corpora*, páginas 82–94. ACL.

Russell, B. 1905. On denoting. *Mind*, (14):479–493.

Solorio, T. 2004. Improvement of named entity tagging by machine learning. Informe técnico, Coordinación de Ciencias Computacionales INAOE (No. CCC-04-004), Puebla, México.

Tjong Kim Sang, E. F. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 1–4, Stroudsburg, PA, USA. ACL.

Tjong Kim Sang, E. F. y J. Veenstra. 1999. Representing text chunks. En *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, páginas 173–179, Stroudsburg, PA, USA. ACL.

Tkachenko, M. y A. Simanovsky. 2012. Named entity recognition: Exploring features. En J. Jancsary, editor, *Proceedings of KONVENS 2012*, páginas 118–127. ÖGAI.

van Dijk, T. A. 1992. *La ciencia del texto*. Paidós, Barcelona.

van Hooland, S., M. D. Wilde, R. Verborgh, T. Steiner, y R. V. de Walle. 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *DSH*, 30(2):262–279.

Wilks, Y. 1998. Sense and texts. *Computational Linguistics and Chinese Language Processing*, 3(2):1–16.