HERRAMIENTAS DE CÓDIGO ABIERTO PARA EL PROCESAMIENTO DEL ESPAÑOL. ANÁLISIS Y COMPARATIVA

Beatriz Valverde Olmedo, Rubén Martínez García
Universidad de Valladolid

Resumen: El procesamiento del lenguaje natural se centra, fundamentalmente, en pretender resolver el problema de la comunicación entre los seres humanos y las máquinas, intentando que las computadoras sean capaces de comprender las relaciones y significados de las palabras, así como todos los mecanismos relacionados con una determinada lengua natural. Actualmente existen múltiples herramientas de dominio público para el procesamiento del lenguaje natural, las cuales típicamente se asocian al inglés. En este trabajo se ha realizado un análisis en profundidad de las librerías y herramientas de código abierto existentes para el procesamiento del español, ya que el español es la segunda lengua materna por número de hablantes, así como la segunda lengua más utilizada en Internet.

Palabras clave: español, etiquetador de categorías gramaticales, herramientas de código abierto, *parser*, procesamiento del lenguaje natural

Abstract: Natural language processing aims to solve the problem of communication between humans and machines, enabling computers capacity for understand the relationships and meanings of words for a specific language, as well as all mechanisms related to a certain natural language. Currently there are multiple public domain tools for natural language processing, typically associated only with English. In this work, we have done an in-depth analysis of the existing open source libraries and tools for the processing of Spanish, because Spanish is the second mother tongue by number of speakers and the second most used language on Internet.

Key words: Spanish language, POS tagger, Open Source toolbox, parser, natural language processing

1. Introducción

A lo largo de toda la historia, el conocimiento se comunica y se maneja, en su mayor parte, a través de las len-

guas naturales. Nuestros días no constituyen una excepción, ya que diariamente utilizamos herramientas y aplicaciones que trabajan con tales lenguas, desde traductores, reconocedores de habla, aplicaciones para la recuperación de información en textos, indexación de contenidos en bases de datos, etc. Asimismo, existe multitud de información en lenguas naturales en Internet y cada día se generan más contenidos con las nuevas formas extendidas de comunicación tanto escrita como hablada, por ejemplo, los repositorios de documentación técnica, aplicaciones de mensajería instantánea, *bots* de comunicación automáticos, reconocedores de voz, todo ello relacionado con el Internet de las cosas (Marr 2016). La interfaz más sencilla para un ser humano de comunicarse con otros seres humanos es mediante una lengua natural, pero para que cualquiera de nosotros podamos utilizar una lengua natural como elemento de comunicación con una máquina (comunicación máquina-humano) es necesario que los computadores sean capaces de entender nuestras afirmaciones, así como el contenido que aparece en las expresiones que usamos. En este instante es donde entran en juego las ciencias de la computación, especialmente la rama de estudio del lenguaje conocida como el procesamiento del lenguaje natural (abreviado PNL o NLP en sus siglas anglosajonas) [1]. El procesamiento del lenguaje natural es un campo de las ciencias de la computación, la inteligencia artificial y la lingüística que se encarga del estudio de la interacción entre el lenguaje humano y los computadores, investigando mecanismos para permitir y facilitar la comunicación entre ambos.

Dentro de la filosofía *Open Source* [2], encontramos diversas herramientas que nos permiten procesar el lenguaje natural de forma automática. Afortunadamente, el código abierto dispone de múltiples bibliotecas de alta calidad para resolver problemas comunes en el procesamiento de textos como análisis de sentimientos, identificación de temas, etiquetado automático de contenidos, etc. Asimismo, existen numerosos estudios sobre las herramientas de procesamiento del lenguaje para lenguas como el inglés (Padró et al. 2010) (Bird et al. 2009), pero es interesante conocer también las herramientas para lenguas tan exten-

didas como el español. En el presente trabajo mostramos un estudio comparativo de las herramientas de código abierto de mayor utilidad existentes dentro del procesamiento del lenguaje natural para el español.

El español es el uno de los idiomas más hablado en el mundo con casi 500 millones de hispanohablantes como lengua materna. Si bien es cierto que, en términos de influencia internacional, el español está por debajo del inglés, tiene un alcance superior al chino y a otras lenguas europeas como el alemán o el francés. El auge del español se siente en todas las áreas y, por ello, también en la red. El inglés y el chino encabezan la clasificación de idiomas más usados en la red, superando ambos más del 20% de la cuota de uso, mientras que el español les sigue con casi un 8%, lo cual se relaciona con la aparición de Internet de forma masiva en nuevas áreas geográficas donde el español es la lengua nativa, como ocurre en diferentes zonas de Hispanoamérica y el Caribe. La situación es similar si la evaluamos desde el punto de vista de las redes sociales, donde el español también ostenta el segundo puesto, solamente superado por el inglés, pero su crecimiento es muy superior frente a la lengua anglosajona. Todos estos datos extraídos del informe *El español: una lengua viva. Informe 2016* del Instituto Cervantes nos proporcionan una idea de la importancia del español en el mundo actual y su gran expansión.

2. El procesamiento del lenguaje natural

El procesamiento del lenguaje natural es una de las ramas más en boga tanto de la inteligencia artificial como de la lingüística computacional. Según el DRAE, *procesar*, en su cuarta acepción, significa `someter datos o materiales a una serie de operaciones programadas', es decir, el NLP somete los elementos y unidades [3] que constituyen las lenguas naturales para reproducir el comportamiento lingüístico de los humanos. Reproducir en un amplio sentido, ya que el NLP no solo se ocupa de interpretar tales lenguas naturales, sino también de comprender, producir y generar las estructuras y mecanismos del lenguaje en general y de las lenguas naturales, como materialización

de esa capacidad, en particular. En definitiva, la comunicación máquina-humano. Aunque los avances en esta disciplina en las últimas décadas son evidentes, debemos ser cautos, como vaticinó Lyons "... está fuera de nuestro alcance saber si algún día el ordenador llegará a simular todos los procesos mentales que intervienen en la producción y comprensión del lenguaje."(227).

Las lenguas naturales son sistemas lingüísticos con una estructura articulada en una serie de niveles interrelacionados, los cuales se componen de un conjunto de elementos y unidades y una serie de reglas relativas a cómo se han de combinar tales elementos y unidades, a saber: nivel fónico, nivel morfológico, nivel sintáctico y nivel semántico. La forma en la que están estructurados estos niveles origina una de las propiedades esenciales de las lenguas, conocida como dualidad de articulación o doble articulación del lenguaje [4]. Esta dualidad posibilita que los elementos finitos y carentes de significado del nivel fónico se combinen para formar unidades con significado de los niveles inmediatamente superiores. En la tabla 1 se recoge un sencillo ejemplo a partir de cuatro fonemas /k/ /a/ /s/ /a/ [5]:

Tabla 1: Ejemplo de niveles de la lengua

/k/ /á/ /s/ /a/	→ cas-a	→ casa	→la casa	→ la casa tiene ventanas
nivel fónico	nivel morfológico		nivel sintáctico	
elementos sin significado	unidades con significado			

Cualquier desarrollo del NLP ha de tener en cuenta estos niveles constitutivos e inherentes de las lenguas naturales, por lo que su aplicación implica una serie de fases o etapas que se corresponden con tales niveles. El *input* [6] puede responder a la modalidad escrita u oral de la lengua, aunque la modalidad oral necesita un módulo que sea capaz de interpretarla y transformarla al lenguaje escrito de tal forma que se asimile a la modalidad escrita para que un computador pueda procesarla. Lo que supone un escollo, a pesar de que las lenguas tienen la propiedad de la transferibilidad del medio:

> ... el isomorfimos completo [entre la modalidad escrita y oral] no es más que un ideal teórico. No existe siste-

> ma ortográfico (al margen de los sistemas de transcripción que los fonetistas han diseñado para este propósito) capaz de representar todas las distinciones importantes del habla (Lyons 13).

Asimismo, es necesario señalar que la fiabilidad de la tecnología actual de las técnicas NLP y las técnicas de ingeniería del conocimiento (KE [7]) no está garantizada en ningún caso para las diferentes fases que componen este tipo de tecnologías. Igualmente, el dominio de aplicación donde se van a utilizar este tipo de técnicas puede causar diversos errores adicionales. Esto implica conocer en profundidad las diferentes alternativas de procesamiento para cada una de las fases y técnicas de procesamiento del lenguaje para poder decidir cuál puede ser más útil en cada dominio de aplicación.

3. Trasfondo tecnológico

Los sistemas de procesamiento del lenguaje natural pueden desarrollarse utilizando diversas técnicas computacionales entre las que destacan la ingeniería del conocimiento y la inteligencia artificial, o bien conjugar ambas técnicas de forma híbrida. En la actualidad, los enfoques más usados son los provenientes del aprendizaje automático o aprendizaje máquina, este se basa en técnicas de inteligencia artificial, puesto que ofrece mejor precisión y requiere menor trabajo manual. De todas formas, los sistemas basados en ingeniería del conocimiento aún resultan útiles en determinados ámbitos como la medicina o textos técnicos, donde se requiere una capacidad para reconocer estructuras complejas.

Se pueden distinguir cuatros fases principales dentro del procesamiento del lenguaje natural (Cohen 2009): preprocesamiento de texto, reconocimiento de nombres de entidades (NER) [8] extracción de contexto y extracción de relaciones.

El preprocesamiento de texto engloba las siguientes fases específicas para las detecciones de oraciones: *tokenizers*, etiquetado gramatical o POS [9] *tagger*, división de palabras en función del corpus del lenguaje o *Treebank* (corpus *parseado*) y analizadores sintácticos o *parsers*. El

etiquetado gramatical se encarga de la identificación de las categorías gramaticales (sustantivos, adjetivos, verbos, etc.) en una sentencia determinada. *Treebank* es un corpus de texto donde cada sentencia que pertenece al corpus asociado tiene una estructura sintáctica añadida a él. OpenNLP (Morton et al. 2005) es el proyecto referencia dentro del procesamiento de texto, o preprocesamiento de texto como lo hemos definido con anterioridad. OpenNLP conjuga una amplia variedad de herramientas para el NLP, entre las que destacan: detección de oraciones, *tokenización,* POS *tagging*, *chunking* y *parsing*, detección de entidades con nombre y correferencias, utilizando técnicas de aprendizaje automático Maxent (Sharma y Dey 2012). En nuestro caso particular para el español, tiene un identificador de oraciones, un *tokenizer* y un POS *tagger*. Otra de las grandes referencias históricas en el procesamiento del lenguaje natural son las herramientas de la Universidad de Stanford, conocidas como Stanford CoreNLP, las cuales incluyen herramientas para el procesamiento del español dentro de sus corpus lingüísticos, destacando por incorporar también el reconocimiento de nombres de entidades. Como la gran mayoría de los trabajos de su época, Stanford CoreNLP provee un análisis sintáctico probabilístico (Briscoe y Carroll 1993) y utiliza algoritmos de programación dinámica para calcular la derivación gramatical más probable de una oración, según un modelo estadístico de la estructura sintáctica de una lengua dada, incluyendo modelos para el español, al igual que ocurría con OpenNLP. Ambos aplican técnicas probabilísticas basadas en algoritmos de Maxent, aunque Stanford CoreNLP también provee modelos de corpus entrenados del español fundamentados en técnicas de perceptrón multicapa.

Los proyectos más influyentes de la última década están centrados, en su mayoría, únicamente en el inglés y apenas se incluyen algunas herramientas para lenguas como el español, lo que requiere la necesidad de un estudio más exhaustivo de las herramientas existentes aparecidas recientemente. En la actualidad, la plataforma líder dentro de las técnicas de preprocesado de textos es NLTK [10] (Bird 2006). Esta incorpora más de cincuenta recursos

léxicos y corpus de conocimiento, incluyendo modelos para el español, con los que se pueden realizar tareas de clasificación, *tokenization, stemming,* POS *tagging, parsing*, así como razonamiento semántico. NLTK posee una amplia comunidad de desarrolladores detrás, encargados de completar y ampliar su gran catálogo de herramientas, así como del desarrollo de libros introductorios al procesamiento del lenguaje natural (Bird, Klein y Loper 2009). Una de las características más importante de NLTK es que permite incorporar herramientas y recursos externos dentro de su elenco de procesos, como por ejemplo el uso de *parsers* y corpus lingüísticos utilizados por la Universidad de Stanford, por lo que es considerada una de las librerías de NLP más completas actualmente dentro del *software* libre.

El reconocimiento de nombres de entidades (NER) es otro de los ejes centrales de las técnicas de NLP. De forma análoga al preprocesado de textos, el sistema más conocido son los modelos utilizados por la Universidad de Stanford en su herramienta Stanford NER (Finkel, Grenager y Manning 2005).

Dentro del proceso de interpretación del lenguaje natural, debemos hablar de las principales ramas de investigación sobre extracción de información, centradas en la desambiguación de entidades (Davis et al. 2012) (van Erp, Rizzo y Troncy 2013), resumen de textos (Becker, Naaman y Gravano 2011), reconocimiento de sentimientos y la minería de opiniones (Maynard, Bontcheva y Rout 2012). En las herramientas de libre distribución para la extracción de información en textos destaca MITIE [11]. Dicha herramienta se centra en el análisis NER y la detección de relaciones binarias, asimismo está preparada para el entrenamiento de extractores y detectores de relaciones más avanzados. MITIE está basada en técnicas avanzadas de aprendizaje máquina, utilizando algoritmos de máquinas de soporte vectorial estructuradas (Joachims, Finley y Yu 2009). Uno de los proyectos más importantes que ha surgido en los últimos años es OpenNER (Agerri et al. 2004), cuyo objetivo principal es proporcionar un conjunto de herramientas listas para realizar diferentes tareas de procesamiento del lenguaje natural, liberadas como *Open*

Source software mediante el uso de servicios web REST, utilizando tecnologías web actuales. OpenNER nace con la necesidad de cubrir una serie de huecos existentes en el flujo de trabajo de cualquier aplicación sobre el análisis de sentimientos y detección de opiniones en textos escritos.

Finalmente, es necesario destacar otro tipo de herramientas dedicadas principalmente a la minería de datos y, en especial, a la minería web, que dentro de sus conjuntos de herramientas integran potentes recursos para el tratamiento de textos. Destaca Pattern (De Smedt y Daelemans 2012), que incorpora a su elenco de técnicas de minería de textos, herramientas para etiquetadores POS, *n-gram search*, análisis de sentimientos, WordNet, algunas de las cuales incorporan corpus lingüísticos del español.

4. Caso de estudio

Para poder evaluar las diferentes técnicas y alternativas dentro de las tecnologías de código abierto para el procesamiento del lenguaje natural, hemos diseñado un diccionario de palabras base con las que construir sentencias y frases complejas. Este diccionario se ha desarrollado a partir de la selección de al menos cinco palabras para cada categoría gramatical capaz de detectar un analizador morfológico, la lista se compone de adjetivos, adverbios, determinantes, preposiciones, pronombres (sujeto y objeto), sustantivos (incluyendo nombres propios y entidades de localizaciones y lugares) y verbos (monoactanciales y biactanciales, incluyendo todos los verbos copulativos). En total, el diccionario/lexicón se compone de más de ochenta palabras y sus posibles derivadas (género y número, plurales y diferentes formas de los verbos elegidos). A partir de nuestro, diccionario se han construido un total de cien sentencias de muestra para poder evaluar de forma detallada las diferentes herramientas del estudio. En términos de extracción de contenidos y reconocimiento de nombres de entidades (NER), el procesamiento de sentencias individuales sin información contextual en las que se encuentran son uno de los grandes retos del NLP. Asimismo, hemos incluido diversos ejemplos entre nuestras sentencias para poder analizar también este tipo de técnicas,

como por ejemplo: *Rubén no ha estado nunca en Nueva York*.

4. 1. Criterios de rendimiento y elección

La evaluación de la calidad de los resultados es un punto importante en la evaluación y comparación de diferentes herramientas, podríamos basarnos en las medidas de precisión (% de éxito) que certifican su calidad de forma genérica, pero en este trabajo hemos tratado de dar una visión particular de cada uno de ellos evaluándolos durante el flujo de trabajo de una aplicación real de procesamiento de textos con una serie de textos seleccionados al azar sobre el lexicón que se ha definido. Las métricas de evaluación que seguiremos se refieren a los estándares en el NLP, establecidas como precisión y exhaustividad (o *recall*). La precisión es una medida que certifica la exactitud de los métodos, identificando la proporción de verdaderos positivos sobre todas las unidades de trabajo. Por su parte, *recall* es una medida más específica, puesto que mide la proporción de verdaderos positivos en la identificación de una determinada clase frente a todas las categorías que deben ser reconocidas. El rendimiento general se controla mediante el valor-F, medida armónica entre la precisión y la exhaustividad.

5. Métodos

Nuestra aproximación pretende evaluar diferentes herramientas y librerías de código abierto que permiten realizar la extracción de entidades y relaciones a partir de sentencias de una lengua natural, así como la identificación de los sentimientos identificados en los textos seleccionados. No se pretenden realizar análisis complejos de relaciones y contenidos de textos al ser un trabajo donde el contexto de los textos elegidos es heterogéneo, simplemente se pretende evaluar la calidad de las herramientas *Open Source* disponibles en la actualidad para componer o aconsejar el mejor flujo de trabajo posible. El límite fijado para el estudio se encuentra en las herramientas de NER y de análisis de sentimientos.

5.1. Extracción de sentencias

La extracción de sentencias es el proceso por el cual a partir de un texto escrito, se extraen las oraciones características. Es un proceso sencillo, aunque pocas *suites* de procesamiento del lenguaje lo incorporan para el español, entre ellas NLTK y OpenNER sí disponen de herramientas, directa o indirectamente, relacionadas con la extracción de oraciones a partir de un texto complejo.

5.2. Preprocesamiento de texto

Las tareas de preprocesamiento de texto orientadas hacia el análisis sintáctico se han llevado a cabo testeando cuatro de las herramientas *Open Source* más populares: OpenNER, OpenNLP, NLTK (configurado con los modelos de StanfordNLP) y Pattern.ES. Las etiquetas del etiquetado gramatical o POS *tagging* con el cual nos encargamos de la identificación de sustantivos, adjetivos, verbos, etc., en una sentencia determinada provienen del corpus lingüístico anotado Penn o Penn TreeBank (Marcus et al. 1994) en todos o bien de una porción del Wiki-Corpus (Reese et al. 2010) en el caso de Pattern.ES para el español (Rocha y Sánchez 2009). Previamente al etiquetado gramatical, es necesario dividir las sentencias en *tokens* o elementos mínimos de contenido que se van a procesar por el etiquetador y realizar un análisis morfológico del contenido en búsqueda de errores como primera fase del procesamiento de sentencias. Adicionalmente, en esta primera fase se puede incluir, si fuera necesario, la eliminación de las palabras vacías (o *stop words* en la jerga anglosajona de NLP). Este trabajo se puede realizar de forma análoga con cualquiera de las herramientas anteriores, en diferentes lenguajes de programación y producen una salida similar, listas de elementos procesables por un analizador sintáctico o por otros tipos de análisis de interés. En esta fase se podría realizar el análisis NER, o identificación de nombres de entidades en el texto, pero hemos preparado un aparte más específico para este tipo de herramientas. Ilustraremos todo el proceso de análisis de herramientas utilizando un ejemplo básico: *los niños comen manzanas*, empezando por el proceso de división de sentencias en *tokens*

o palabras que sirvan como fuente para el procesamiento morfológico.

- Sentencia: *los niños comen manzanas*.

- Resultado:
  [*los*]
  [*niños*]
  [*comen*]
  [*manzanas*]

El análisis morfológico incluye la categorización de palabras, así como la identificación de las raíces de cada una de ellas, raíces morfológicas o lemas, comúnmente denominada lematización o *stemming*. El *stemming* consiste en la eliminación de los sufijos para obtener su forma canónica o raíz de la palabra y es necesario para múltiples procesos de búsqueda e indexación de contenidos en texto. Existen algunos algoritmos de *stemming* que ayudan en sistemas de recuperación de información. *Stemming* aumenta el *recall* (Powers 2007) que es una medida sobre el número de documentos que se pueden encontrar con una consulta. Igualmente, la lematización resulta parecida, pero incluye un pequeño matiz, consistente en la eliminación de los sufijos y el reemplazo de las letras de las palabras para obtener su forma canónica o raíz de la palabra, obteniendo el lema que se acepta como representante de la misma en una determinada lengua, es decir, la entrada que encontraríamos en un diccionario.

Dentro de las herramientas que hemos discutido, las más útiles en este tipo de técnicas son NLTK para *stemming* que utiliza el sistema SnowBall, un pequeño lenguaje para procesamiento del lenguaje o bien una librería específica en Python [12], denominada PyStemmer. Por su parte, dentro de las herramientas más útiles para la lematización se encuentra Pattern.es, es una las pocas herramientas de *software* libre que implementan la obtención de las formas normales de los verbos, así como sus conjugaciones. Seguimos con nuestro ejemplo base con el que demostrar esta nueva técnica de procesamiento:

- Entrada:
  [*los*]
  [*niños*]
  [*comen*]
  [*manzanas*]

- Salida:
  [*el*]
  [*niño*]
  [*comer*]
  [*manzana*]

La categorización/etiquetado de palabras o POS *tagger* es la fase en la cual se identifican las etiquetas sintácticas de cada palabra de una sentencia (sustantivos, determinantes, verbos, etc). Todas las herramientas que hemos identificado anteriormente son útiles en esta tarea: Pattern.ES, OpenNLP, NLTK y OpenNER, además utilizan las mismas técnicas, asignan el etiquetado y desambiguan de forma sincronizada e incluyen lematización, aunque de forma interna en muchos casos. Todos los analizadores morfológicos tienen un diccionario de palabras y corpus de conocimiento como hemos indicado anteriormente, para la lengua española el diccionario Freeling (Padró y Stanilovsky 2012) incorpora el 90% de esta lengua y contiene herramientas para realizar análisis de textos, entre ellas diversas herramientas de etiquetado de palabras. A continuación, observamos el funcionamiento de esta técnica a través de nuestro ejemplo ilustrativo:

- Entrada:
  [*el*]
  [*niño*]
  [*comer*]
  [*manzana*]

- Salida [13]:
  [*el*]: [DT]
  [*niño*]: [NN]
  [*comer*]: [VB]
  [*manzana*]: [NN]

También se puede realizar el etiquetado de sentencias a través de la sentencia completa como hace la herramienta Pattern.ES, obteniendo el siguiente resultado, donde el lematizador y el análisis de categorías se realizan de forma atomizada:

- Salida:

  [*los*]: /DT/B-NP/O/el
  [*niños*] /NNS/I-NP/O/niño
  [*comen*] /VB/B-VP/O/comer
  [*manzanas*]: /NNS/B-NP/O/manzana

Realizado el análisis morfológico, puede ser de interés disponer de correctores ortográficos libres, uno de los más conocidos dentro del área de *software* libre es MySpell, el corrector ortográfico utilizado por OpenOffice.

El siguiente paso es el análisis de la estructura de cada sentencia, este análisis también se denomina análisis sintáctico. La entrada de un analizador sintáctico es una lista de *tokens* provenientes del analizador léxico y su objetivo es comprobar que la estructura sintáctica sea válida de acuerdo con una gramática dada, en nuestro caso la gramática del español. El analizador se encarga de asignar una estructura a sentencias que previamente carecían de orden. Todos los analizadores sintácticos evaluados (NLTK, OpenNLP y OpenNER) se basan en analizadores estadísticos en función de corpus previamente anotados. Desgraciadamente, la mayoría de corpus anotados son de pago, salvo algunas excepciones como el Penn Treebank del que hemos hablado anteriormente. Pocos de ellos disponen de soporte para el español, o bien están directamente creados para este propósito. Este tipo de análisis se divide en tres fases: 1) crear el corpus de entrenamiento, 2) método de inferencia, normalmente utiliza gramáticas libres de contexto probabilísticas (Collins 2011) y 3) herramientas de análisis que se encargan de crear un árbol de decisión según las probabilidades de cada palabra de ser un determinado tipo sintáctico de acuerdo con su categoría gramatical. Observamos a continuación los resultados obtenidos a través de nuestro ejemplo básico:

- Entrada:
  [*los*]: /DT/B-NP/O/el
  [*niños*] /NNS/I-NP/O/niño
  [*comen*] /VB/B-VP/O/comer
  [*manzanas*]: /NNS/B-NP/O/manzana
- Salida obtenida: en los gráficos 1 y 2 se ilustra el análisis sintáctico del procesamiento del lenguaje natural en español.

Gráfico 1: Ejemplo de árbol sintáctico (Freeling)

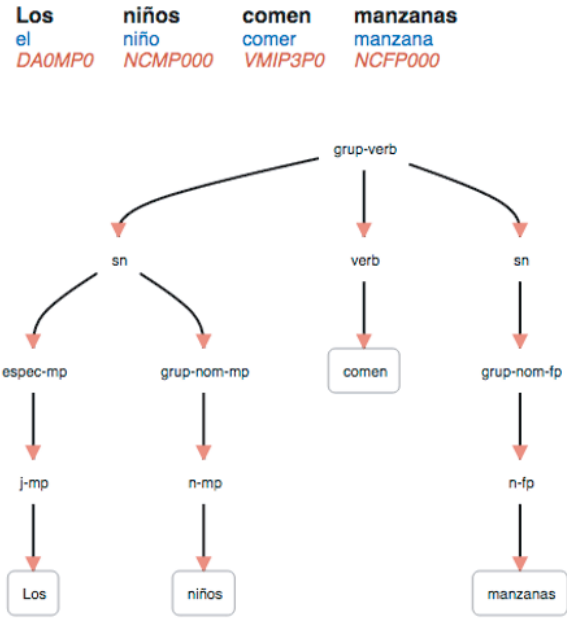
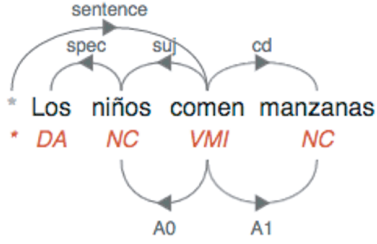


Gráfico 2: Ejemplo de relaciones sintácticas (Freeling)



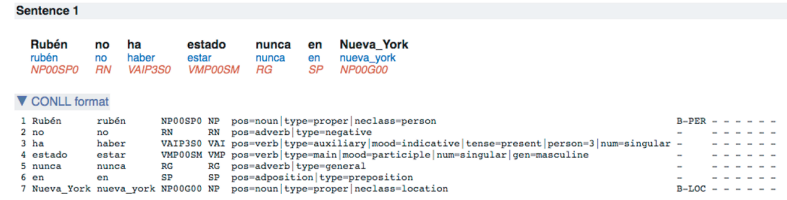
Finalmente podemos hablar del análisis de grupos lingüísticos o *chunks*, por ejemplo [*el gato*][*come*][*pescado*], son herramientas internas que se utilizan en los analizadores sintácticos, pero que pueden ser útiles en otros ámbitos de forma independiente y no todas las librerías analizadas las soportan. Dentro de las herramientas evaluadas, solamente Freeling dispone de este tipo de recursos de forma nativa.

5.3. Reconocimiento de nombres de entidades

El análisis NER consiste en la localización de los nombres de organizaciones, localizaciones y personas (en ocasiones se puede añadir alguna categoría adicional como 'miscelánea', o bien fechas, datos de tiempo y datos numéricos) dentro de un texto. El reconocimiento de entidades en textos se divide en dos tareas, la primera tarea se ocupa de localizar en el texto aquellas secuencias de palabras que puedan constituir una entidad diferenciada, mientras que la segunda tarea está dedicada a la clasificación de este tipo de entidades ya localizadas. OpenNER y la herramienta de la Universidad de Stanford, Stanford NER, disponen de recursos que se encargan de realizar ambas tareas de forma conjunta, tanto la localización de elementos o entidades individuales como la clasificación de las mismas en personas, localización u organización. Por el contrario, NLTK no dispone de ambas funcionalidades, únicamente nos permite obtener la lista de entidades diferenciales. El análisis NER lo vamos a ilustrar con la herramienta Freeling, pero con una sentencia que incluye nombres específicos para esta técnica:

- Entrada: *Rubén nunca ha estado en Nueva York.*
- Salida: obtenemos la identificación de las palabras *Rubén* como persona, y *Nueva York* como localización, podemos verlo en el gráfico 3 del análisis.

Gráfico 3: Análisis NER (Freeling)



5.4. Análisis de sentimientos

Una de las múltiples tareas que engloba el procesamiento del lenguaje natural es la clasificación de textos, que consiste en la categorización del contenido de un texto en función de la búsqueda de relaciones y palabras clave que permitan englobarlo en una o varias categorías previamente definidas. Existen multitud de textos que son de carácter subjetivo, donde el procesamiento no debería limitarse únicamente a la clasificación genérica de los mismos. El análisis de sentimientos permite paliar esta deficiencia, ayudando a clasificar los documentos en función de la polaridad de opinión que presenta su autor. Estas técnicas combinan técnicas de NLP y minería de textos, donde tiene un soporte básico en la inteligencia artificial y los métodos de clasificación de contenidos. Dentro de las herramientas que hemos evaluado en el presente trabajo, únicamente OpenNER realizar esta tarea de forma autónoma sin ayuda externa de otros recursos.

6. Resultados

De forma conjunta al listado de las herramientas más influyentes en procesamiento del lenguaje natural de código abierto, es necesario medir la precisión para determinar el comportamiento de cada una de ellas en las diferentes técnicas de NLP que se han aplicado durante este trabajo. También resulta imprescindible realizar una pequeña crítica sobre la facilidad de uso y su aplicabilidad en cada una de las técnicas aplicadas.

6.1. Preprocesamiento de textos

Las herramientas de etiquetado de categorías certifican unos resultados del 96,2% para los modelos de lexicón uti-

lizados en el español con Penn Treebank, frente a unos resultados cercanos al 90% para el procesamiento utilizando WikiCorpus. Se expresan los resultados obtenidos en la tabla 2:

Tabla 2: Resultados de precisión de los analizadores sintácticos

Nota: F1-value = 2 (precisión**recall*) / (precisión+*recall*)	
NLTK (Stanford)	Precisión 87% y Recall 96% y F1-value 91%
NLTK (WikiCorpus)	Precisión 82% y Recall 90% y F1-value 85%
OpenNLP	Precisión 83% y Recall 94% y F1-value 86%
OpenNER	Precisión 89% y Recall 96% y F1-value 92%

Evaluar las discrepancias con tan baja desviación en contextos dispares y con un *dataset* del tamaño que hemos dispuesto en este trabajo no arroja resultados concluyentes sobre cuál de los objetos de estudio es más útil. Desde el punto de vista de sencillez de uso, las herramientas basadas en servicios web (OpenNER) y en lenguajes interpretados como Python (NLTK) son las que permiten un acceso y gestión de información mucho más rápido y sencillo para no expertos en programación, facilitando la aproximación a herramientas de procesamiento del lenguaje natural a más público que otro tipo de librerías como pueden ser OpenNLP o las librerías de Stanford.

6.2. Reconocimiento de nombres de entidades

NLTK expone una precisión aproximada del 96% en su etiquetador de entidades, donde solamente prevé la identificación de sentencias o entidades con un significado especial. Nuestro *dataset* no está lo suficientemente orientado hacia el contraste de este tipo de análisis de texto, por lo que hemos obtenido el 100% de éxito tanto con NLTK como con OpenNER a la hora del proceso de etiquetado. Desde el punto de vista de profundidad y riqueza de resultados, OpenNER otorga al usuario una interfaz más sencilla de trabajo gracias al uso de servicios web para acceder a este tipo de recursos, así como una herramienta más potente ya que, como hablábamos en la sección anterior, no solamente permite la búsqueda de entidades sino también su clasificación.

6.3. Análisis de sentimientos

El análisis de sentimientos es una medida subjetiva, no permite un cálculo efectivo de la precisión de los métodos, sino que, simplemente, podemos interpretar si sus resultados son correctos de forma subjetiva por parte de un experto. En nuestro caso, hemos evaluado el desempeño y la calidad de los resultados propuestos por las herramientas de análisis de sentimientos de OpenNER, obteniendo resultados altamente satisfactorios, puesto que OpenNER ofrece al usuario una forma rápida, limpia y sencilla de obtener el análisis de sentimientos de un texto.

6.4. Resumen

En la tabla 3, y a modo de resumen, se recogen todas las técnicas analizadas en el presente artículo, así como las herramientas de código libre que pueden utilizarse para cada una de ellas.

Tabla 3: Resumen de las librerías de código abierto y técnicas analizadas

	NLTK	OpenNER	Stanford	Pattern.es	Freeling	Otros
Tokenizer	x	x	X	x	x	OpenNLP
Stemmer	x	x	X	x	x	OpenNLP PyStemmer
POS tagger	x	x	X	x	x	OpenNLP
Tree tagger	x	x	X		x	OpenNLP
Análisis sintáctico	x		X		x	
Análisis de chunks					x	
Análisis NER	Parcial	x	Parcial		x	
Análisis de sentimientos		x				

7. Conclusiones

El procesamiento del lenguaje natural es el núcleo central de muchas aplicaciones y utilidades que se requieren diariamente: traducción automática, clasificación e indexación de textos, extracción de información de textos, corrección automática, compiladores, intérpretes de contenidos, etc., pero todo esto es solo el principio, puesto que el futuro de la tecnología radica en la interacción directa entre humanos y computadores a través de la lengua natural, tanto escrita como principalmente hablada, y el procesamiento del lenguaje natural es la herramienta clave para poder llevar a cabo este salto.

En este trabajo hemos pretendido recopilar las herramientas de mayor interés para desarrollar cualquier trabajo dentro de iniciativas de código abierto que impliquen algunas de las áreas de mayor utilidad dentro del procesamiento del lenguaje natural.

El amplio abanico de herramientas abiertas que hemos detallado en este trabajo no hace más que reforzar la teoría de que existen suficientes recursos que permiten extraer información de textos de forma automática accesibles a un gran público de expertos, no solo en materia de computación, sino también a expertos lingüistas.

BIBLIOGRAFÍA

AGERRI, Rodrigo, Montse Cuadros, Sean Gaines y German Rigau. "OpeNER: Open Polarity Enhanced Named Entity Recognition". *Procesamiento del Lenguaje Natural* 51 (2013). Impreso.

BECKER, H., M. Naaman, and L. Gravano. "Selecting Quality Twitter Content for Events". *Proceedings of the Fifth International Conference on Weblogs and Social Media* (2011). Impreso.

BIRD, Steven. "NLTK: The Natural Language Toolkit". *Proceedings of the COLING/ACL on Interactive presentation sessions* (2006): 69-72. Impreso.

BIRD, Steven, Ewan Klein y Edward Loper. *Natural language processing with Python: Analyzing Text with the Natural Language Toolkit*. O'reilly, 2009. Impreso.

BRISCOE, Ted y John CARROLL. "General Probabilistic LR parsing of natural language (Corpora) with unification-based grammars". *Computal Linguistics-Special issue on using large corpora: I* 19. 1 (March 1993): 25-29. Impreso.

COHEN, Kevin Bretonnel y Dina DEMNER-FUSHMAN. *Biomedical Natural Language Processing*. John Benjamins Publishing Company, 2014. Impreso.

COLLINS, Michael. *Probabilistic (lexicalized) Context Free Grammars*. Course of University of Columbia, 2011.

DAVIS, A., A. VELOSO, A. SOARES, A. LAENDER y W. MEIRA Jr. "Named entity disambiguation in streaming data". *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island, Korea (July 2012): 815–824. Impreso.

DE SMEDT, T. y W. DAELEMANS. "Pattern for Python". *Journal of Machine Learning Research.* 13 (2012) 2031–2035. Impreso.

FINKEL, Jenny Rose, Trond GRENAGER y Christopher MANNING. "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling". *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (2005): 363-370. Impreso.

Instituto Cervantes. *El español: una lengua viva. Informe 2016*. Madrid, 2016.

JOACHIMS, T., T. FINLEY y Chun-NAM YU. "Cutting-Plane Training of Structural SVMs". *Machine Learning* 77. 1 (2009): 27-59. Impreso.

LYONS, John. *Una introducción al lenguaje y a la lingüística.* Barcelona: Teide, 1984. Impreso.

MARCUS, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz y Britta Schasberger. "The Penn Treebank: annotating predicate

argument structure". *Proceedings of the workshop on Human Language Technology* (1994): 114-119. Impreso.

MARR, Bernard. "Why natural language processing will change everything". *DataInformed: Big Data and Analytics in the Enterprise*, agosto 2016, http://data-informed.com/why-natural-language-processing-will-change-everything/. Último acceso enero de 2017.

MARTINET, André. *Elementos de lingüística general.* Madrid: Gredos, 1974. Impreso.

MAYNARD, D., K. BONTCHEVA, y D. ROUT. "Challenges in developing opinion mining tools for social media". *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC*. Turkey (2012). Impreso.

MORTON, Thomas, Joern KOTTMANN, Jason BALDRIDGE y Gann BIERNER. *Opennlp: A java-based nlp toolkit*, 2010, https://opennlp.apache.org. Último acceso enero de 2017.

PADRÓ, Lluís, Miquel COLLADO, Samuel REESE, Marina LLOBERES e Irene Castellón. "FreeLing 2.1: Five Years of Open-Source Language Processing Tools. A: International Conference on Language Resources and Evaluation". *7th International Conference on Language Resources and Evaluation*. La Valetta: 2010.

PADRÓ, Lluís y Evgeny STANILOVSKY. "FreeLing 3.0: Towards Wider Multilinguality". *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.* Istanbul, Turkey (May 2012). Impreso.

Powers, David M. W. "Evaluation: From precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation". *Technical Report SIE-07-001.* University of South Australia (December 2007). Impreso.

REESE, Samuel, Gemma BOLEDA, Montse CUADROS, Lluís PADRÓ y German RIGAU. "Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus". *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10).* La Valleta, Malta (May 2010). Impreso.

ROCHA, Marta Alicia y Joan Andreu SÁNCHEZ. "Machine Translation of the Penn Treebank to Spanish". *Proceedings of the I Iberian SLTech* (2009). Impreso.

SHARMA, Anuj y Shubhamoy DEY. "A comparative study of feature selection and machine learning techniques for sentyment analysis". *Proceedings of the 2012 ACM Research in Applied Computation Symposium.* San Antonio, Texas (October 23-26 2012): 1-7. Impreso.

VAN ERP, M., G. RIZZO y R. TRONCY. "Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning". *Proceedings of the 3rd Workshop on Making Sense of Microposts* (2013). Impreso.

NOTAS

[1] En este trabajo utilizamos las siglas en inglés *Natural Language Processing* por ser de uso más común.

[2] Open Source o código abierto, es la filosofía mediante la cual el *software* es distribuido y desarrollado libremente.

[3] Seguimos la diferenciación terminológica de Lyons entre *elemento* y *unidad* (Lyons 9).

[4] Término acuñado por Martinet (1974).

[5] El lector puede acudir al DRAE para comprobar que ninguno de los fonemas recogidos posee significado.

[6] Omitimos la recomendación ofrecida sobre el anglicismo *input* por el DPD, por ser este término más usado en informática que sus equivalentes españoles *entrada* o *introducción de datos.*

[7] *Knowledge Engineering.*

[8] Utilizamos las siglas anglosajonas NER, *Named-entity recognition.*

[9] *Part of Speech.*

[10] *Natutal Language ToolKit.*

[11] *MIT Information Exchange.*

[12] Lenguaje de programación interpretado, muy común en el NLP.

[13] El etiquetado gramatical está dado por el etiquetado estándar definido por Stanford University.