Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 23-30 recibido 03-04-2017 revisado 11-05-2017 aceptado 18-05-2017

Propuesta de un sistema de clasificación de entidades basado en perfiles e independiente del dominio

Proposal for a domain independent named entity classification system based on profiles

Isabel Moreno
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
imoreno@dlsi.ua.es

M.Teresa Romá-Ferri
Dpt. Enf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
mtr.ferri@ua.es

Paloma Moreda
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
moreda@dlsi.ua.es

Resumen: El reconocimiento y la clasificación de entidades nombradas (RCEN) es clave para muchas aplicaciones de procesamiento de lenguaje natural. Sin embargo, la adaptación de un sistema RCEN resulta costosa, ya que la mayoría solo funcionan adecuadamente en el dominio para el que fueron desarrollados. Considerando esta premisa, se evalúa si un sistema de clasificación de entidades nombradas basado en perfiles y aprendizaje automático obtiene los mismos resultados independientemente del dominio del corpus de entrenamiento. Para ello, hemos experimentado con 6 tipos de entidades de dos dominios en español: general y médico. Aplicando técnicas para equilibrar la distribución de las clases, se ha logrado que la diferencia de F1 entre ambos dominios sea de 0,02 (F1: 50,36 versus 50,38, respectivamente). Lo cual apoya la independencia del dominio del sistema basado en perfiles.
Palabras clave: Clasificación de entidades nombradas, Perfiles, Aprendizaje automático, Dominio independiente, Español, Corpus desequilibrados

Abstract: Named Entity Recognition and Classification (NERC) is a prerequisite to other natural language processing applications. Nevertheless, the adaptation of NERC systems is expensive given that most of them only work appropiately on the domain for which they were created. Bearing this idea in mind, a named entity classification system, which is profile and machine learning based, is evaluated to determine if the results are maintained regardeless of the domain of the training corpus. To that end, it is tested on 6 types of entities from two different domains in Spanish: general and medical. Applying techniques to balance class distribution, the difference in terms of F1 between domains is 0.02 points (F1: 50.36 versus 50.38, respectively). These results support the domain independence of our profile-based system.
Keywords: Named entity classification, Profiles, Machine learning, Domain independent, Spanish, Imbalanced corpora

1 Introducción

El Reconocimiento y la Clasificación de Entidades Nombradas (RCEN) tiene dos objetivos. Primero, identificar las menciones de nombres propios en un texto, conocida como la fase de reconocimiento (REN). Segundo, asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas, llamada fase de clasificación (CEN) (Marrero et al., 2013). Ambos objetivos pueden abordarse de manera conjunta o separada.

Los sistemas RCEN juegan un papel importante en muchas aplicaciones que procesan información textual. Por un lado, el RCEN es un prerequisito para diversas tareas como minería de opiniones (Marrero et al., 2013) o generación de lenguaje natural (Vicente y Lloret, 2016). Por otro lado, existe un efecto positivo en el rendimiento de estas aplicaciones al incluir un sistema RCEN, como en el caso de la generación de resúmenes (Fuentes y Rodríguez, 2002; Alcón y Lloret, 2015).

La mayoría de sistemas RCEN solo funcionan adecuadamente en el dominio para el que fueron desarrollados. Cada dominio suele tener requisitos característicos y, por lo tanto, diferentes tipos de entidades con los que tra-

ISSN 1135-5948

bajar. Como resultado, estos sistemas están diseñados ad-hoc para un conjunto reducido de entidades predefinidas. Por ello, se requiere un esfuerzo para adaptar una herramienta RCEN a un nuevo dominio, que cuente con restricciones diferentes y un conjunto de entidades propio (Marrero et al., 2013).

Considerando los actuales antecedentes, nuestro propósito final es desarrollar un sistema RCEN independiente del dominio, que afronte el problema, con dos módulos separados, REN y CEN, secuenciales.

En este trabajo nos centraremos en el desarrollo del módulo CEN suponiendo un módulo REN perfecto, que evita cualquier sesgo potencial. Dicho módulo CEN empleará perfiles y aprendizaje automático supervisado. El CEN desarrollado se caracterizará por ser independiente del dominio, es decir, mantendrá sus resultados a pesar de cambiar el dominio del corpus de entrenamiento y el conjunto de entidades predefinidas a clasificar.

Para confirmar esta independencia del dominio, el módulo CEN se evaluará con dos corpus de dominios diferentes en español: (a) noticias del dominio general (Tjong Kim Sang, 2002) y (b) fichas técnicas de medicamento del dominio médico (Moreno, Moreda, y Romá-Ferri, 2012).

En los siguientes apartados mostramos las características de sistemas RCEN independientes del dominio y sistemas CEN (Sección 2). En la sección 3, detallamos nuestro módulo CEN, así como los materiales, la experimentación y la discusión. Terminamos con las conclusiones y el trabajo futuro en la sección 4.

2 Antecedentes: Sistemas RCEN

En las dos últimas décadas, muchas investigaciones se han centrado en RCEN (Marrero et al., 2013). No obstante, las aproximaciones son difíciles de reutilizar, ya que la mayoría se centran en un solo dominio. Así, en diversas competiciones internacionales podemos encontrar sistemas RCEN desarrollados para un único dominio, por ejemplo, el general (Tjong Kim Sang, 2002; Sang y De Meulder, 2003; Màrquez et al., 2007; Ji, Nothman, y Hachey, 2015) o el médico (Uzuner, Solti, y Cadag, 2010; Segura-Bedmar, Martınez, y Herrero-Zazo, 2013; Pradhan et al., 2014).

Sin embargo, son pocos los estudios sobre RCEN que se declaran independientes del dominio. Tkachenko y Simanovsky (2012) experimentaron con varios géneros textuales presentes en el corpus OntoNotes. Su propuesta obtiene una F1 que oscila entre 50 y el 75 %. Kitoogo y Baryamureeba (2008) definen un sistema RCEN que se probó en dos dominios (general y legislativo). En concreto, se experimentó entrenando en el dominio general (Sang y De Meulder, 2003) y evaluando en el legislativo, y viceversa. Su aproximación alcanzó una F1 próxima al 92 % y 70 %, respectivamente.

Ahora bien, si nos centramos exclusivamente en la CEN, no encontramos sistemas que hayan probado su capacidad en diferentes dominios. Si bien algunos son evaluados con corpus diferentes, el dominio no cambia, y, sin embargo, los resultados se ven afectados de forma negativa. Por ejemplo, en el trabajo de Gamallo et al. (2014), son probados Freeling (Carreras, Marquez, y Padró, 2002), OpenNLP[1] y CitiusNEC (Gamallo et al., 2014) con los corpus Hetero y CoNLL2002. Ambos corpus, de dominio general, recogen noticias y, en el caso de Hetero, también entradas de la Wikipedia. En los dos primeros casos, las pruebas realizadas muestran diferencias sustanciales en el valor de F1 para ambos corpus (OpenNLP - 79,02 % versus 65,65 %; Freeling - 75,98 % versus 65,56 %); mientras que esa diferencia es mucho menor en el último caso (CitiusNEC - 66,89 % versus 66,40 %).

Los resultados de estos antecedentes muestran que los sistemas no han mostrado un rendimiento óptimo en diferentes dominios o géneros textuales. Cuando cambia el corpus, aún manteniendo el dominio, se observa un detrimento importante en las prestaciones de la mayoría de sistemas.

Este trabajo propone un sistema CEN basado en perfiles y aprendizaje automático, que mantenga sus resultados aunque cambie el dominio y el conjunto de entidades.

3 Propuesta de clasificación de entidades basada en perfiles

En esta sección, describiremos un sistemas CEN basado en aprendizaje automático y perfiles, así como los requisitos del cambio de dominio (Sección 3.1). Después, caracterizaremos tanto los corpus sobre los que experimentaremos (Secciones 3.2 y 3.3) como las

[1] `https://opennlp.apache.org` (Último acceso: 1/Junio/2017)

medidas de evaluación (Sección 3.4), y, finalmente, mostraremos los resultados (Sección 3.5) y su discusión (Sección 3.6).

3.1 Sistema de clasificación de entidades nombradas

El CEN propuesto está basado en perfiles y se refiere a una colección de descriptores únicos, que no son más que un conjunto de conceptos relevantes encontrados en un corpus. Este trabajo es una adaptación del método de Lopes y Vieira (2015) pero se diferencia del nuestro en dos sentidos. Por un lado, Lopes y Vieira (2015) utilizan conceptos para generar perfiles, mientras que nosotros usamos lemas de palabras con significado[2]. Por otro lado, aquí calculamos la similitud entre entidades y perfiles mediante aprendizaje supervisado, pero Lopes y Vieira (2015) definen su propia medida de similitud.

Nuestro sistema consta de dos fases:

La primera fase, llamada *Generación de perfiles*, entrena el sistema y se divide, a su vez, en 5 pasos: (I) Se analiza el corpus, previamente anotado con entidades nombradas, tanto para separar el texto en oraciones y *tokens*, como para obtener el lema y la categoría gramatical. (II) El corpus se divide en dos conjuntos: positivo (+), instancias de la entidad objetivo; y negativo (−), instancias del resto de clases. (III) Para cada uno de estos conjuntos se extraen los lemas de palabras con significado en ventanas de tamaño V[3]. Por limitaciones de espacio, para cada entidad solo se muestra la V con los mejores resultados. V se ha determinado empíricamente entre 20 y 40. En este punto, los lemas que solo acompañan a esta entidad constituyen la lista principal de descriptores; pero si aparecen también con otras entidades, forman parte de la lista común de descriptores. (IV) A cada uno de los elementos de estas listas se les asigna unos índices de relevancia, basados en el TFDCF (Lopes y Vieira, 2015). En este paso, se generan los perfiles que son pares de lemas-relevancia para la lista principal y la común:$\{lema(i),\ relevancia(lema(i))\}$. El tamaño de perfil (T) será la suma del número de pares descriptor-índice de ambas listas. Por restricciones de espacio, solo mostramos las 3 mejores: 2000[4], 1000[5] y 50[6]. (V) Y, finalmente, para cada entidad, se entrena su propio clasificador usando el algoritmo de aprendizaje supervisado *Voted Perceptron* (Freund y Schapire, 1999) implementado en Weka (Hall et al., 2009), donde las características de cada modelo son los perfiles. El valor de cada característica es su índice de relevancia.

Finalmente, la fase de *Aplicación de perfiles* es la encargada de clasificar las entidades reconocidas por un REN. En este trabajo se asume un análisis lingüístico realizado con Freeling (Padró y Stanilovsky, 2012) y la salida de un REN perfecto, aunque estos pueden cambiarse por cualquier otro. Para cada una de las entidades reconocidas se generan nuevos perfiles, con las mismas restricciones que en la fase de generación. Después, se comparan los perfiles generados en este paso con los generados anteriormente y se calcula la similitud. El sistema determina la clase final en base al perfil más similar.

El CEN propuesto no precisa ninguna modificación en el sistema para su aplicación a un nuevo dominio. Este proceso es directo. Basta con generar los perfiles para las nuevas clases a partir del nuevo corpus de entrenamiento y el conjunto de tipos de entidades con el que trabajar. Con estos datos, el sistema ya es capaz de generar los nuevos perfiles con los que volver a entrenar el CEN, consiguiéndose así un CEN para un dominio diferente.

3.2 Corpus

Se han empleado dos corpus de dos dominios diferentes (general y médico), que determinan el conjunto de etiquetas con el que trabajar.

El corpus *CoNLL2002* (Tjong Kim Sang, 2002) es una colección de artículos en Español de la agencia de noticias EFE. Contiene cuatro tipos de entidades: persona, organización, localización y miscelánea. Nosotros descartaremos esta última por no tener una aplicación práctica real, como sugiere Marrero et al. (2013). Este corpus se divide en tres conjuntos: entrenamiento (18797 entidades), desarrollo (4351 entidades) y evaluación (3558 entidades). El modelo de aprendizaje automático es inferido del conjunto de entrenamiento

[2]Son sustantivos, verbos, adjetivos y adverbios.

[3]$\frac{V}{2}$ palabras antes y después de la entidad.

[4]Cada lista contiene hasta 1000 descriptores.

[5]Cada lista contiene hasta 500 descriptores.

[6]La lista principal contiene un máximo de 50 descriptores pero la común está vacía, para comprobar si esta última es necesaria.

y la evaluación aquí presentada se realiza en el conjunto de evaluación.

El gold standard *DrugSemantics* (Moreno, Moreda, y Romá-Ferri, 2012) es una colección de 5 Fichas Técnicas de Medicamento (FTM) en español. Contiene 780 oraciones y más de 2000 entidades anotadas manualmente. En este trabajo usamos las tres clases con mayor frecuencia en el corpus: proceso clínico (724 entidades), principio activo (657 entidades) y unidad de medida (557 entidades). La baja frecuencia de los tipos restantes no permite emplear aprendizaje automático. La evaluación se realiza mediante validación cruzada de 5 iteraciones, es decir, en cada iteración se entrena con 4 FTM y se evalúa con la restante para, finalmente, obtener como resultado la media de todas las iteraciones.

3.3 Corpus equilibrados

Nuestra metodología divide los corpus de entrenamiento en dos (positivo versus negativo) para generar perfiles. Esto provoca que la distribución de ambas clases sea más desequilibrada que la inicial. Por ejemplo, la partición de entrenamiento del corpus CoNLL2002 contiene sólo un 23 % de entidades tipo Persona (positiva), mientras que la clase negativa representa el 77 % restante. Similar es el caso del corpus DrugSemantics, por ejemplo, las ocurrencias de tipo Unidad de Medida (positiva) suponen únicamente el 25 %, mientras que las negativas engloban el 75 % restante.

Este desequilibrio en los corpus de entrenamiento es muy habitual. No obstante, conduce frecuentemente a que los algoritmos tradicionales de aprendizaje automático supervisado resulten sesgados hacia la clase negativa (también mayoritaria) y, por eso, la clase positiva (o minoritaria) sale perjudicada, a pesar de que usualmente contiene los datos de mayor interés (López et al., 2012).

En los últimos años, se han propuesto diversos mecanismos para afrontar el desequilibrio entre clases (López et al., 2012):

(a) Mecanismos de muestreo: Alteran la distribución de las clases en los datos de entrenamiento. Existen 3 opciones: (I) sub–muestreo (*under-sampling*), elimina instancias generalmente de la clase mayoritaria; (II) sobre-muestreo (*over-sampling*), añade instancias nuevas o las replica; o (III) una mezcla de ambos.

(b) Algoritmos sensitivos al coste: Minimizan el coste de las clasificaciones incorrectas, asumiendo que los costes de error son diferentes para cada clase. Existen tres opciones: (I) crear nuevos algoritmos; (II) introducir un pre-proceso a algoritmos existentes que modifique los datos de entrenamiento, por ejemplo asignando pesos a las instancias de acuerdo al coste de los errores; o (III) incluir un postproceso a los algoritmos tradicionales que altere el umbral de clasificación del clasificador.

A pesar de que existen estudios sobre el comportamiento de estas técnicas, no es posible extraer conclusiones respecto a qué mecanismo es el más adecuado. López et al. (2012) indican que "ambas aproximaciones son buenas y equivalentes". Además, Wei y Dunbrack (2013) advierten que estos mecanismos dependen tanto de la cantidad de datos como del problema.

Por ello, en este trabajo haremos uso de tres técnicas de equilibrio, implementadas en Weka (Hall et al., 2009): (a) *sub-muestreo*, igualando el número de instancias en ambas clases aleatoriamente, debido a su sencillez y poco coste computacional; (b) *sobre-muestreo*, mediante la técnica SMOTE (Chawla et al., 2002), puesto que la generación de nuevas instancias de la clase minoritaria ha dado buenos resultados en otros RCEN (Tomanek y Hahn, 2009; Al-Rfou et al., 2014); y (c) clasificación sensitiva al coste, mediante un pre–proceso que asigna pesos a las instancias del corpus de entrenamiento[7], ya que en pruebas iniciales hemos observado que ofrece mejores resultados que el uso del post-proceso para los mismos costes.

3.4 Medidas de evaluación

Nuestra aproximación se evaluará para las 6 entidades empleando las medidas tradicionales de Precisión, Cobertura y la medida F1 tanto para la clase positiva (+) como la negativa (-). Además, calcularemos una media ponderada para agrupar los resultados de ambas clases (negativa y positiva) de acuerdo a su distribución. Finalmente, se ofrecerá una macro-media global del sistema para cada corpus que será la media aritmética de los

[7] En Weka empleamos la clase CostSensitiveClassifier con la siguiente matriz de coste: [0 1; 5 0]. Dichos costes se escogieron aleatoriamente pero garantizando que el coste de los errores en la clase positiva fuera mayor que en la negativa (5>1).

E(|V|)	|T|	Prec+	Cob+	F1+	Prec-	Cob-	F1-	$\overline{Prec}$	$\overline{Cob}$	$\overline{F1}$
O (|20|)	50	53,69	15,57	24,14	62,50	91,29	74,20	61,50	34,30	67,12
	1000	50,64	36,50	42,42	65,12	76,92	70,53	61,02	57,43	60,34
	2000	**51,32**	**38,93**	**44,27**	65,75	76,04	70,52	61,44	58,95	60,54
P (|20|)	50	42,05	5,03	8,99	79,88	98,19	88,10	78,95	14,93	82,53
	1000	46,39	10,48	17,09	80,60	96,85	87,98	79,01	26,95	79,21
	2000	**43,41**	**10,75**	**17,23**	80,57	96,35	87,76	78,67	28,42	77,89
L (|20|)	50	50,00	4,89	8,91	70,13	97,86	81,71	69,53	13,19	78,01
	1000	**52,26**	**24,54**	**33,40**	73,17	90,18	80,79	70,18	48,76	69,65
	2000	53,11	23,62	32,69	73,08	90,86	81,01	70,38	47,40	70,34
Macro-media		48,99	24,74	31,63	73,16	87,52	79,69	70,10	45,38	69,36

Abreviaturas (por orden de aparición) : E, entidad; |V|, Tamaño de ventana; |T|, tamaño del perfil; +, clase positiva; Prec+, Precisión+ ; Cob+, Cobertura+; F1+, medida $F_{\beta=1}$;-, clase negativa; Prec-, Precisión-; Cob-, Cobertura-; F1+, medida $F_{\beta=1}$; $\overline{Prec}$, Precisión ponderada; $\overline{Cob}$, Cobertura ponderada; $\overline{F1}$, $F_{\beta=1}$ ponderada; O, Organización; P, Persona; L, Localización

Tabla 1: Resultados con el corpus CoNLL2002

E(|V|)	|T|	Prec+	Cob+	F1+	Prec-	Cob-	F1-	$\overline{Prec}$	$\overline{Cob}$	$\overline{F1}$
PC (|20|)	50	60,70	46,24	51,81	75,27	85,17	79,61	71,90	71,75	70,81
	1000	**59,04**	**59,33**	**58,14**	78,40	78,01	77,69	73,20	71,42	71,61
	2000	59,75	53,08	55,28	77,06	81,58	78,86	72,48	71,83	71,47
PA (|40|)	50	40,05	24,79	29,21	77,57	87,55	82,02	69,43	72,10	69,54
	1000	**39,13**	**45,03**	**39,40**	81,01	77,70	78,62	72,12	69,01	69,52
	2000	36,32	40,37	36,48	80,12	76,43	77,75	70,76	67,34	68,26
UM (|40|)	50	40,36	15,72	21,67	79,38	94,22	85,88	72,40	76,88	72,34
	1000	**45,05**	**18,11**	**23,88**	79,45	93,69	85,76	73,30	76,87	72,74
	2000	47,76	16,19	23,59	79,54	94,91	86,34	74,22	77,64	73,31
Macro-Media		47,74	40,82	40,47	79,62	83,13	80,69	72,87	72,43	71,29

Abreviaturas (por orden de aparición) : E, entidad; |V|, Tamaño de ventana; |T|, tamaño del perfil; +, clase positiva; Prec+, Precisión+ ; Cob+, Cobertura+; F1+, medida $F_{\beta=1}$;-, clase negativa; Prec-, Precisión-; Cob-, Cobertura-; F1+, medida $F_{\beta=1}$; $\overline{Prec}$, Precisión ponderada; $\overline{Cob}$, Cobertura ponderada; $\overline{F1}$, $F_{\beta=1}$ ponderada; PC, Proceso Clínico; PA, Principio Activo; UM, Unidad de Medida

Tabla 2: Resultados con el corpus DrugSemantics

mejores resultados de cada clase.

3.5 Resultados

Las Tablas 1 y 2 recogen una comparativa de los resultados de cada dominio, general y médico, respectivamente. Las líneas marcadas en negrita muestran la mejor F1 de la clase positiva (F1+) para cada tipo de entidad.

Existe una gran diferencia entre los resultados de la clase positiva y los de la negativa, independientemente de la entidad o el dominio (fila Macro-media). Se puede observar que los modelos están sesgados hacia la clase negativa porque producen una cobertura excesivamente baja en la clase positiva. En concreto, Cob+ es menor del 45 % como norma general. Esto implica que la F1+ también sea excesivamente baja en 5 de las 6 entidades. Por el contrario, la F1- siempre es alta (mayor del 70 %). Por esta razón los resultados ponderados son adecuados (mayor del 60 %) en todas las entidades, pero menores que los negativos.

Para afrontar dicho sesgo, las Tablas 3 y 4 recogen, de manera global y para cada entidad, una comparativa entre el mejor resultado con y sin equilibrio para la clase positiva en los dominios general y médico, respectivamente. De nuevo, las líneas en negrita destacan la mejor F1+ en cada tipo de entidad.

En el corpus CoNLL2002 (Tabla 3), observamos que el sub-muestreo (u) ofrece la mejor F1+ en Organización y Localización. Sin embargo, SMOTE (sobre-muestreo - o) logra los mejores resultados para Persona. Destacar

que la clasificación sensitiva al coste no aparece en la tabla pero siempre es la segunda mejor opción, existiendo una diferencia pequeña con respecto a la mejor solución (entre 0,05 y 5,33 puntos).

En lo referente al corpus DrugSemantics (Tabla 4), se aprecia que el sobre-muestreo (o) consigue la mejor F1+ en Proceso Clínico y Principio Activo, aunque con escasa diferencia respecto al sub-muestreo (u). Por el contrario, Unidad de Medida consigue un mayor resultado directamente con sub-muestreo (u). En este corpus, los resultados de la clasificación sensitiva al coste siempre son superados por las técnicas de muestreo.

Los resultados de las Tablas 3 y 4 revelan que estas técnicas han permitido mejorar la F1+ en ambos dominios. La mejora siempre conlleva un incremento de cobertura, con una ligera reducción en la precisión en algunos casos. En concreto, en el dominio general el porcentaje de mejora media de F1+ es mayor del 109,62 % (% de Mejora en la Tabla 3) y en el dominio farmacológico es de un 31 % (% de Mejora en la Tabla 4).

Respecto a la cobertura, las Tablas 3 y 4 muestran una mejora importante en ambos corpus al incluir mecanismos de muestreo, en comparación con los bajos resultados que se obtenían previamente (Tablas 1 y 2). El corpus DrugSemantics mejora en un 75,7 % de media. El caso del corpus CoNLL2002 es especialmente llamativo, ya que la cobertura aumenta más de un 180 % de media.

En cuanto a la precisión, la mejora no siempre es positiva al introducir las técnicas de muestreo. En el caso del corpus CoNLL2002 la precisión mejora casi un 20,50 % de media, aunque empeora ligeramente en Localización (en menos de 10 puntos). El descenso medio en la precisión en DrugSemantics es algo mayor y supone una pérdida media de alrededor del 9 %, ya que sólo Principio Activo mejora la precisión en casi un 2 %.

Estos resultados nos llevan a dos conclusiones. (1) Las técnicas de equilibrio son necesarias para el desarrollo de un sistema independiente del dominio basado en perfiles y aprendizaje automático binario, atendiendo a los corpus utilizados. (2) Cada tipo de entidad requiere un clasificador propio ya que sus necesidades (mecanismo de equilibrio, tamaño de ventana y de perfil) son diferentes.

E(|V|,|T|)	**A**	**Pr+**	**Co+**	**F1+**
O(|20|,|2000|)	S	51,32	38,93	44,27
	u	**51,85**	**65,14**	**57,74**
P(|20|,|50|)	S	42,05	5,03	8,99
	o	**65,67**	**30,66**	**41,81**
L(|20|,|2000|)	S	53,11	23,62	32,69
	u	**43,95**	**62,27**	**51,53**
Macro-Media	S	44,58	18,74	24,02
	e	53,72	52,69	50,36
Mejora		9,14	33,95	26,33
% de Mejora		20,50	181,17	109,62

Abreviaturas (por orden de aparición): E, entidad; |V|, tamaño de ventana; |T|, tamaño del perfil; A, Aproximación; +, clase positiva; Pr+, Precisión+; F1+, medida $F_{\beta=1}$+; Co+, Cobertura+; O, Organización; s, sistema Sin equilibrar; u, sUb-muestreo; P, Persona; o, sObre-muestreo; L, Localización; e, sistema Equilibrado

Tabla 3: CoNLL2002 con y sin equilibrar

E(|V|,|T|)	**A**	**Pr+**	**Co+**	**F1+**
PC(|20|,|2000|)	S	59,75	53,08	55,28
	o	**53,19**	**66,64**	**58,82**
PA(|40|,|2000|)	S	36,32	40,37	36,48
	o	**38,16**	**71,35**	**47,95**
UM(|40|,|2000|)	S	47,76	16,19	23,59
	u	**39,04**	**54,59**	**44,38**
Macro-Media	S	47,94	36,55	38,45
	e	43,46	64,20	50,38
Mejora		-4,48	27,65	11,93
% de Mejora		-9,34	75,70	31,00

Abreviaturas (por orden de aparición): E, entidad; |V|, tamaño de ventana; |T|, tamaño del perfil; A, Aproximación; +, clase positiva; Pr+, Precisión+; F1+, medida $F_{\beta=1}$+; Co+, Cobertura+; PC, Proceso Clínico; s, sistema Sin equilibrar; PA, Principio Activo; UM, Unidad de Medida; u, sUb-muestreo; o, sObre-muestreo; e, sistema Equilibrado

Tabla 4: DrugSemantics con y sin equilibrar

3.6 Discusión

Nuestros resultados han mostrado que el CEN basado en perfiles es independiente del dominio. Aplicando técnicas de equilibrio, se ha logrado una diferencia global media de F1+ entre dominios de 0,02 puntos (Tablas 3 y 4 fila Macro-media e: 50,36 versus 50,38).

En cuanto a la comparación de nuestro sistema con los de la Sección 2, tanto los

RCEN declarados independientes de dominio como los CEN, esta no está exenta de limitaciones. Por un lado, muchas veces los corpus utilizados son diferentes; por ejemplo, CoNLL2002 (Tjong Kim Sang, 2002) en español y CoNLL2003 (Sang y De Meulder, 2003) en inglés. Por otro lado, mientras que en este trabajo nos centramos únicamente en la fase CEN, los restantes sistemas proporcionan resultados conjuntos de ambas fases. Además, el conjunto de entidades a clasificar no siempre es el mismo; por ejemplo, aquí no se ha considerado la clase miscelánea. No obstante, es necesario analizar si los resultados obtenidos están en consonancia.

Respecto a la comparación con sistemas RCEN declarados independientes de dominio, la Tabla 5 recoge la diferencia entre los mejores y peores resultados de F1+ global de ambos RCEN: DINERS (Kitoogo y Baryamureeba, 2008) y TKSIM (Tkachenko y Simanovsky, 2012). Se observa que ambos presentan una diferencia mayor de 20 puntos en términos de F1+, mientras que en nuestro caso es significativamente menor: 0,02 puntos.

Respecto a la comparación con los sistemas CEN en español, la Tabla 6 recoge la F1+ global en dichos CEN con CoNLL2002 y Hetero. Estos se comparan con nuestro sistema en el corpus CoNLL2002 y DrugSemantics. Además, la diferencia entre los dos corpus es mostrada en la columna Dif. Se observa, que si bien los valores de F1 obtenidos son algo inferiores al resto de sistemas, esta diferencia es mucho menor cuando la comparación se hace con un corpus diferente al empleado en su desarrollo (columna C2 frente C1). Respecto a las consecuencias de cambio de dominio, la aproximación basada en perfiles mantiene sus resultados, mientras que el resto de sistemas reducen sus resultados entre 0,49 y 14 puntos.

Sistema	**Dif**	**C1**	**C2**
PerfilesE	**0.02**	**50, 36$^{\$}$**	**50, 38$^{\$\$}$**
TKSIM	>25	$< 50^{\%}$	$< 75^{\%\%}$
DINERS	>20	70*	92**

Abreviaturas (por orden de aparición): Dif, Diferencia absoluta entre mejor y peor F1+; C1, mejor F1+; C2, peor F1+; $^{\$}$, corpus CoNLL2002; $^{\$\$}$, corpus DrugSemantics; $^{\%}$, corpus OntoNotes-bc-msnbc; $^{\%\%}$, corpus OntoNotes-mz-sinorama; *, corpus CoNLL2003; **, corpus "Uganda Courts of Judicature".

Tabla 5: Comparación con RCEN independientes del dominio

CEN	**Dif**	**C1**	**C2**
PerfilesE	**0,02**	**50,36**	**50,38**
CitiusNEC	0,49	66,89	66,40
Freeling	10,42	75,98	65,56
OpenNLP	13,37	79,02	65,65

Abreviaturas (por orden de aparición): Dif, Diferencia absoluta entre mejor y peor F1+; C1: F1+ global en corpus CoNLL2002; y C2: F1+ global en corpus DrugSemantics o Hetero.

Tabla 6: Comparativa con CEN

4 Conclusiones y trabajo futuro

En este artículo hemos presentado un sistema de clasificación de entidades nombradas basado en perfiles y aprendizaje automático independiente del dominio. Para confirmar dicha independencia, hemos experimentado con 6 tipos de entidades en dos corpus de dominios diferentes: 3 tipos de entidades del dominio general (CoNLL2002) y 3 del médico (DrugSemantics).

Sin necesidad de adaptar el sistema CEN basado en perfiles de un dominio a otro, y utilizando técnicas de muestreo (sub-muestreo y sobre-muestreo), se ha mostrado su efectividad para la clasificación independientemente del dominio a partir de los resultados de F1+. En concreto, la diferencia al cambiar de dominio es de 0,02 puntos (F1+ Macro-media general: 50,36 versus médico: 50,38).

Los resultados son prometedores, pero nuestro CEN es un trabajo en progreso y necesita continuar mejorando. El trabajo futuro se centrará en tres objetivos. (1) Se orientará a mejorar los resultados de la clasificación, incluyendo nuevas características independientes del dominio que son frecuentemente usadas en sistemas RCEN (como los afijos). (2) Se continuará analizando el desequilibro en los corpus en dos líneas. Primero, se planteará probar otras estrategias para dividir los corpus en positivo y negativo. Segundo, se estudiará la presencia de otras dificultades asociadas comúnmente al desequilibrio, como el ruido o el solapamiento entre clases. (3) Se evaluará el funcionamiento en más dominios.

Agradecimientos

Investigación financiada por el Gobierno de España (TIN2015-65100-R; TIN2015-65136-C02-2-R) y la Generalitat Valenciana (PROMETEOII/2014/001).

Bibliografía

Al-Rfou, R., V. Kulkarni, B. Perozzi, y S. Skiena. 2014. POLYGLOT-NER: Massive Multilingual Named Entity Recognition. *ArXiv e-prints*, (October).

Alcón, Ó. y E. Lloret. 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües. *Linguamática*, 7(1):53–63, Julio.

Carreras, X., L. Marquez, y L. Padró. 2002. Named entity extraction using adaboost. En *Proceeding of the 6th Conference on Natural Language Learning*.

Chawla, N. V., K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Freund, Y. y R. E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Fuentes, M. y H. Rodríguez. 2002. Using cohesive properties of text for automatic summarization. *Jotri'02*.

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.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Ji, H., J. Nothman, y B. Hachey. 2015. Overview of TAC-KBP2015 Entity Discovery and Linking Tasks. En *Proceedings of Text Analysis Conference 2015*.

Kitoogo, F. y V. Baryamureeba. 2008. Towards domain independent named entity recognition. En *Strengthening the Role of ICT in Development*, volumen IV. Fountain publishers, páginas 84 – 95.

Lopes, L. y R. Vieira. 2015. Building and Applying Profiles Through Term Extraction. En *X Brazilian Symposium in Information and Human Language Technology*, páginas 91–100.

López, V., A. Fernández, J. G. Moreno-Torres, y F. Herrera. 2012. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608.

Màrquez, L., L. Villarejo, M. A. Martí, y M. Taulé. 2007. SemEval-2007 Task 09 : Multilevel Semantic Annotation of Catalan and Spanish. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, páginas 42–47.

Marrero, M., J. Urbano, S. Sánchez-Cuadrado, J. Morato, y J. M. Gómez-Berbís. 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5):482–489.

Moreno, I., P. Moreda, y M. Romá-Ferri. 2012. Reconocimiento de entidades nombradas en dominios restringidos. En *Actas del III Workshop en Tecnologías de la Informática*. páginas 41–57.

Padró, L. y E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. En *Proceedings of the Language Resources and Evaluation Conference*, páginas 2473–2479.

Pradhan, S., N. Elhadad, W. W. Chapman, S. Manandhar, y G. Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. páginas 54–62.

Sang, E. F. T. K. y F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. En *Proceedings of the 7th Conference on Natural Language Learning*, páginas 142–147.

Segura-Bedmar, I., P. Martınez, y M. Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). En *Proceedings of the 7th International Workshop on Semantic Evaluation*, páginas 341–350.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task. En *Proceeding of the 6th Conference on Natural Language Learning*.

Tkachenko, M. y A. Simanovsky. 2012. Selecting Features for Domain-Independent Named Entity Recognition. En *Proceedings of KONVENS 2012*, páginas 248–253.

Tomanek, K. y U. Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. En *Proceedings of the fifth international conference on Knowledge capture*, páginas 105–112.

Uzuner, O., I. Solti, y E. Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–8.

Vicente, M. y E. Lloret. 2016. Exploring Flexibility in Natural Language Generation throughout Discursive Analysis of New Textual Genres. En *Proceedings of the 2nd International Workshop FETLT*, Sevilla, Spain.

Wei, Q. y R. L. Dunbrack. 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, 8(7).