Reconocimiento de entidades nombradas para el idioma español utilizando *Conditional Random Fields* con características no supervisadas

Jenny Linet Copara Zea

Orientador: Dr. José Eduardo Ochoa Luna

Jurado:
Dr. Camilo Thorne – University of Stuttgart – Alemania
Dr. Fabio Cozman – Universidade de Sao Paulo – Brasil
Dr. Yván Túpac Valdivia – Universidad Católica San Pablo – Perú
Dr. Alex Cuadros – Universidad Católica San Pablo – Perú

Tesis presentada al
Centro de Investigación e Innovación en Ciencia de la Computación (RICS)
como parte de los requisitos para obtener el grado de
Maestro en Ciencia de la Computación.

Universidad Católica San Pablo – UCSP
Marzo de 2017 – Arequipa – Perú

A Dios, por todo lo que me ha dado, a mis padres por su apoyo incondicional.

Abreviaturas

NLP *Natural Language Processing*

NER *Named Entity Recognition*

CRF *Conditional Random Field*

HMM *Hidden Markov Model*

CoNLL *Conference on Natural Language Learning*

RNN *Recurrent Neural Network*

CNN *Convolutional Neural Network*

SBW *Spanish Billion Words*

Agradecimientos

Agradezco a Dios por sus bendiciones y guía.

Agradezco a mi familia, por su apoyo y soporte.

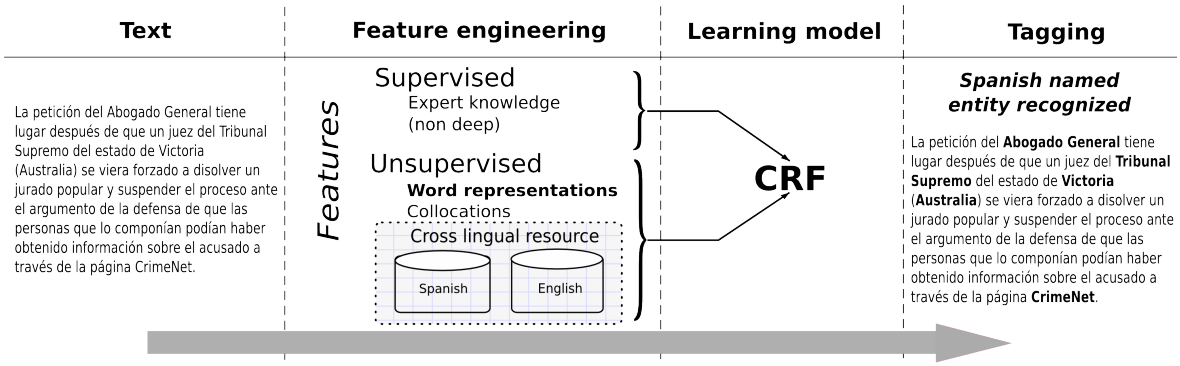
Agradezco a la Universidad Católica San Pablo por brindarnos la oportunidad de ser parte del Programa de Maestría.

Agradezco al Programa de Maestría de la Universidad Católica San Pablo por la oportunidad de estudiar en esta casa de estudios. Asimismo, deseo agradecer de manera especial al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico e Innovación Tecnológica (FONDECYT-CIENCIACTIVA), que mediante Convenio de Gestión UCSP-FONDECYT Nº 011-2013, han permitido la subvención y financiamiento de mis estudios de Maestría en Ciencia de la Computación en la Universidad Católica San Pablo (UCSP).

Agradezco de forma muy especial a mi asesor Dr. José Eduardo Ochoa Luna por su dedicación y guía en la elaboración de esta tesis.

Agradezco al Grupo Data & Web Science de la Universidad de Mannheim (Alemania), mediante el Prof. Heiner Stuckenschmidt y Prof. Simone Ponzetto. De la misma manera, agradezco al Dr. Goran Glavaš por su tiempo y valiosas discusiones. De forma muy especial al Dr. Camilo Thorne por su guía, tiempo y esfuerzo en la elaboración de esta tesis.

Abstract



Named entity recognition (NER) is an important task in Natural Language Processing, which identify entities in texts in a specific language. Several researchers have focused mainly on English language.

Recent research in this field for English language have shown that using unsupervised features such as word embeddings can boost NER. In this work we investigate whether unsupervised features can boost supervised NER in Spanish language. To do so, we used unsupervised features through word embeddings and collocations as additional features in a classifier *Conditional Random Field* (CRF). Experimental results (82.44 % F-score in CoNLL-2002) show that our proposal is comparable with some state of the art approaches for Spanish language, in particular when we used *cross-lingual word representations*.

Keywords: Natural Language Processing, NER for Spanish, Conditional Random Fields, Unsupervised features, Word representations, Word embeddings, Collocations.

Resumen

El reconocimiento de entidades nombradas es una tarea relevante en el área de Procesamiento de Lenguaje Natural, su función es identificar entidades en textos para un idioma dado. El estudio de esta tarea se ha enfocado principalmente en el idioma inglés.

Recientes estudios en el idioma inglés han mostrado que utilizar características no supervisadas tales como *word embeddings* mejoran el reconocimiento de entidades nombradas. En este trabajo se investiga si características no supervisadas pueden mejorar la tarea de NER supervisado en el idioma español. Para esto, se propone utilizar características no supervisadas mediante *word representations* y colocaciones, así como características adicionales en un clasificador *Conditional Random Field* (CRF). Resultados experimentales (82.44 % de F-score en el corpus CoNLL-2002) muestran que el enfoque propuesto, en particular cuando se utiliza *cross-lingual word representations*, es comparable a abordajes de aprendizaje profundo, actualmente el estado del arte para NER en español.

Palabras clave: Procesamiento de lenguaje natural, NER para español, *Conditional Random Fields*, Características no supervisadas, *Word representations*, *Word embeddings*, Colocaciones.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

La comunicación es parte de la vida cotidiana, es la forma en que las personas se expresan, a través del habla, gestos, sonidos, escritura (por medios digitales o manuscritos), entre otros. Asimismo, es posible obtener información mediante su procesamiento, proceso que los humanos han llevado a cabo casi sin darse cuenta, pero cuando se tiene una considerable cantidad de datos a procesar resulta más conveniente aplicar un modelo computacional para realizar determinada tarea (Ponce Gallegos et al., 2014). Es así que el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés *Natural Language Processing*) toma por referencia el lenguaje humano para aprender modelos computacionales útiles para determinadas tareas (Jurafsky y Martin, 2009).

Diferentes tareas en *Natural Language Processing* (NLP) requieren del uso de sub-tareas que realicen un procesamiento previo, como en el caso de traducción automática, recuperación de información, verificación gramatical y clasificación de textos (Nivre, 2000; Çelik, 2012) que requieren del uso de un etiquetador *Part-of-speech* para conocer la categoría gramatical de cada palabra (Manning y Schütze, 1999). Otro claro ejemplo y foco principal de este trabajo de tesis, es la tarea de reconocer entidades nombradas (NER por sus siglas en inglés *Named Entity Recognition*), entendiéndose por entidades los nombres propios de las personas, lugares, direcciones, ciudades, países, entre otros. De la misma forma, NER ha sido utilizado como parte de las tareas de resumen automático (para extraer el significado de textos), recuperación de información, traducción automática, *Question Answering*, minería de textos (Szarvas, 2008), *Entity Linking* (Stern et al., 2012), entre otros.

Diversos modelos de aprendizaje máquina empleados para NER han sido exitosos, en particular el modelo *Conditional Random Fields* (CRF), el cual al considerar características como entradas (observaciones), establece relaciones con sus respectivas etiquetas (Lafferty et al., 2001). En ese contexto, la búsqueda y elección de las características (*feature engineering*) requiere de especial atención. Asimismo, todo el proceso requiere de conocimiento previo, ya sea acerca del tema (dominio del texto) o rasgos del idioma en el que se está aplicando la tarea, es decir, conocimiento supervisado por expertos lo cual hace que este proceso sea supervisado. No obstante, existen datos am-

pliamente disponibles que no han sido procesados por expertos (datos sin etiquetar) que a su vez son útiles para aprender características de forma no supervisada.

1.1. Motivación y Contexto

En NER en el idioma inglés se han explorado diferentes enfoques de aprendizaje, contrario a lo ocurrido en el español. Tras las propuestas de modelamiento del lenguaje[1] con mayor grado de representación de lenguaje, surgió la propuesta del uso de características no supervisadas en un modelo CRF para NER en inglés. Entonces, se emplearon *clusters* de palabras y *"word embeddings"*. En ambos casos se alcanzó resultados competitivos con el estado del arte.

Por otro lado, NER en el idioma español fue estudiado en el desafío de la VI Conferencia en Aprendizaje Natural de Lenguaje Computacional (CoNLL), donde Carreras et al. (Carreras et al., 2002) alcanzaron un valor de 81.39 % en la métrica *F-measure*. Carreras et al. (Carreras et al., 2002) exploraron un conjunto definido de características y recursos supervisados. Luego, en esta tesis, se ha seleccionado un conjunto de características (supervisadas y sin el uso de recursos externos) para el idioma español logrando alcanzar un *F-measure* de 80.02 %.

Finalmente, como sugiere la propuesta de NER para el idioma inglés del uso de características no supervisadas en un modelo CRF, es posible que, de la misma forma se alcancen resultados competitivos para NER en español.

1.2. Planteamiento del Problema

Dado que el modelo CRF requiere de un conjunto de características como entrada, es necesario realizar un proceso de búsqueda y elección de características (*feature engineering*) el cual requiere de conocimiento del dominio, lo que conlleva a emplear tiempo y esfuerzo de un experto para hacerlo.

El proceso de *feature engineering* se ha llevado cabo de forma supervisada (Tjong Kim Sang, 2002; Tjong Kim Sang y De Meulder, 2003; Faruqui y Padó, 2010; Bingel y Haider, 2014), mediante la definición de un conjunto de características dependientes del dominio e idioma, y a su vez de recursos externos supervisados (Carreras et al., 2002; Ratinov y Roth, 2009; Turian et al., 2010; Passos et al., 2014) (lista de palabras comúnmente conocidas en cada entidad). Por ejemplo, la propuesta con mejores resultados en CoNLL para el español alcanzó un 79.28 % de *F-measure* con características supervisadas y un 81.39 % de *F-measure* con conocimiento externo, notando un considerable incremento en *F-measure* al añadir recursos externos lo cual muestra su dependencia.

[1] Modelamiento del lenguaje también conocido como *language modeling*.

En este trabajo, para NER en el idioma español se realiza *feature engineering* explorando características aprendidas de forma no supervisada en el idioma español e inglés, para reducir la dependencia de recursos supervisados.

En consecuencia, el enfoque abordado para este trabajo se basa en emplear CRF para el reconocimiento de entidades nombradas en el idioma español utilizando un conjunto de características (principalmente no supervisadas), aprendidas por algoritmos de *word representations* y recursos lingüísticos en la etapa de *feature engineering*.

1.3. Objetivos

Mejorar la tarea de clasificación de NER en el idioma español al emplear características no supervisadas en el modelo gráfico probabilístico *Conditional Random Fields*.

Asimismo, se definen los siguiente objetivos específicos:

- Identificar el conjunto de características a ser consideradas como línea base para el idioma español.
- Investigar y proponer el uso de características no supervisadas en CRF que mejoren el reconocimiento de entidades nombradas en español.
- Determinar si el uso de características no supervisadas multilenguaje mejora la tarea de clasificación de NER en el idioma español.
- Realizar un análisis comparativo con las actuales técnicas en el estado del arte para NER en español.

1.4. Contribuciones

Las principales contribuciones de esta tesis se resumen en:

- Investigación del impacto de *word representations* (dado por *clustering* de palabras y *word embeddings*) como características en un modelo CRF para el idioma español (Copara et al., 2016b). Cabe señalar que la aplicación de *word representations* como características en el modelo CRF para NER en el idioma español es un abordaje nuevo. Esta investigación llevó a la publicación el 6.º Workshop de Entidades Nombradas (NEWS) en la Conferencia de la Asociación en Lingüística Computacional (ACL) 2016.
- Análisis exploratorio y comparativo del uso de características no supervisadas (colocaciones de palabras y *word representations*) en un modelo CRF evaluado en

los corpora [2] CoNLL-2002 (Copara et al., 2016a) y AnCora (Copara et al., 2016c) (ambas evaluaciones sobre el idioma español) para la tarea de reconocimiento de entidades. Esta investigación llevó a la publicación en 15.ª Conferencia Iberoamericana en Inteligencia Artificial (IBERAMIA) 2016 y en la 5.ª Conferencia Brasileña en Sistemas Inteligentes (BRACIS) 2016.

- Combinación de *word representations* de diferentes idiomas en NER para el idioma español alcanzando los mejores resultados en el corpus CoNLL-2002 (Copara et al., 2016b) y AnCora (Copara et al., 2016c).
- Una alternativa económica y competitiva a las técnicas que actualmente son el estado del arte para NER en el idioma español (Copara et al., 2016b), basadas en el aprendizaje profundo, que demandan un mayor tiempo en entrenamiento y sofistificación.

1.5. Organización de la tesis

El documento está organizado en 6 capítulos. En el Capítulo 2 se presentan los conceptos necesarios para sustentar la propuesta. Los tópicos abordados son: Procesamiento de Lenguaje Natural, *Named Entity Recognition*, clasificación de secuencias y la definición de características no supervisadas.

En el Capítulo 3 se abordan los trabajos relacionados al reconocimiento de entidades nombradas en el idioma español y al uso de características no supervisadas en CRF. Luego, en el Capítulo 4 se presenta la propuesta de la tesis empleando CRF como modelo de aprendizaje con un conjunto de características supervisadas y características adicionales no supervisadas.

En el Capítulo 5 se presenta la metodología de evaluación, detalles de implementación y resultados alcanzados con la propuesta de la tesis. Se explica los conjuntos de datos empleados y las métricas de evaluación, así como también el análisis de los resultados. Finalmente en el Capítulo 6 se presentan las conclusiones de la tesis.

[2] Plural de corpus.

Capítulo 2

Conceptos Previos

En este capítulo se presentan conceptos necesarios para sustentar la propuesta. Primero, se define NLP y NER, luego la clasificación de secuencias en texto. Finalmente, características en secuencias de textos.

2.1. Procesamiento de Lenguaje Natural

Procesamiento de Lenguaje Natural o *Natural Language Processing* (también llamado Lingüística Computacional) es el área de Inteligencia Artificial que permite el estudio y análisis de tareas específicas en un idioma dado (Kumar, 2008; Reese, 2015). Las tareas de NLP intervienen cuando es necesario obtener información de los datos presentes en el lenguaje natural, cuando realizar procesamiento manual es muy costoso o inaceptable. A continuación se explica brevemente (Reese, 2015) algunas tareas de NLP:

- *Machine Translation:* Involucra la traducción de un lenguaje natural a otro.
- *Summarization:* Resumen de colecciones de texto, documentos, artículos, párrafos.
- *Sentiment Analysis:* Esta tarea tiene por objetivo determinar las actitudes, sentimientos o gustos de las personas.
- *Semantic Role Labeling:* Asignación de roles semánticos como *Agent*, *Theme* o *Instrument*, para frases en una oración (Carreras y Màrquez, 2005).
- *Part-of-Speech Tagging*: En esta tarea se asigna las diferentes etiquetas gramaticales a cada palabra de la oración, tal como sustantivo, verbo, entre otros.
- *Named Entity Recognition(NER)*: Involucra extraer entidades nombradas en el texto, tal como ubicaciones, personas, organizaciones, entre otros (Tjong Kim Sang, 2002; Tjong Kim Sang y De Meulder, 2003).

2.2. Named Entity Recognition (NER)

Una entidad nombrada (en inglés *named entity*) es una secuencia de palabras que designa una entidad en el mundo real (Aggarwal y Zhai, 2012), por ejemplo "Perú", "Apple Inc.", "Banco de Crédito del Perú". Una entidad nombrada también puede ser entendida como aquella que puede ser referenciada con un nombre propio (Lin, 2011).

La tarea de reconocimiento de entidades nombradas comúnmente denominada como NER (por *Named Entity Recognition*) consiste en identificar las entidades del texto y clasificarlas en un conjunto predefinido de tipos, tales como persona, organización y ubicación (Aggarwal y Zhai, 2012).

A continuación en el Cuadro 2.1 se presenta un ejemplo de etiquetado NER. La oración de ejemplo es: *"Backus, La Ibérica y BCP permanecen más de un siglo en el mercado local."* . Donde ORG hace referencia a la entidad *"Organización"*, AMOUNT a la entidad *"Cantidad"* y NE indica que la palabra no es una entidad.

Backus	,	La	Ibérica	y	BCP	permanecen	más
ORG	NE	ORG		NE	ORG	NE	NE
de	un	siglo	en	el	mercado	local	.
NE	AMOUNT		NE	NE	NE	NE	NE

Cuadro 2.1: Oración de ejemplo de etiquetado NER

Es importante señalar que NER se ha empleado en etapas iniciales de tareas como: *Question Answering*, *Machine Translation*, *Information Retrieval*, *Text Mining* y *Automatic Summarization* (Szarvas, 2008). Durante *Automatic Summarization* se empleó en la extracción del significado de textos, para identificar ideas principales.

2.2.1. Estilo de anotación

El estilo de la anotación es otro aspecto relevante a tomar en cuenta, es decir la forma en la que se ha realizado el etiquetado de cada elemento[1]. Tomando como referencia el ejemplo del Cuadro 2.1, con mayor atención en las entidades "La Ibérica" y "un siglo" (ambas con más de una palabra) con un estilo de anotación directo (utilizando el nombre de la entidad en cada elemento que la compone).

Los estilos de anotación comúnmente empleados en la tarea son:

- **IO** (Tjong Kim Sang y De Meulder, 2003): Diferenciando entre la anotación dentro de una entidad (***I**nside*) y fuera de ella (***O**utside*)

[1]Los elementos en el texto son palabras, signos de puntuación, números y símbolos.

$$w_i = \begin{cases} I - entidad, & si\ w_i\ es\ entidad, \\ O, & no\ entidad. \end{cases}$$

- **BIO** (Tjong Kim Sang, 2002): Anotando el inicio de la entidad (***B****eginning*), dentro (***I****nside*) y fuera (***O****utside*) de ella.

$$w_i = \begin{cases} B - entidad, & si\ w_i\ es\ inicio\ de\ entidad, \\ I - entidad, & si\ w_i\ es \text{ continuación } de\ entidad, \\ O, & no\ entidad. \end{cases}$$

- **BILOU** (Ratinov y Roth, 2009): Anotando el inicio de la entidad (***B****eginning*), dentro (***I****nside)*, fin(***L****ast*) y fuera (***O****utside*) de ella. También cuando la entidad se compone de un solo elemento (***U****nique*).

$$w_i = \begin{cases} B - entidad, & si\ w_i\ es\ inicio\ de\ entidad, \\ I - entidad, & si\ w_i\ es \text{ continuación } de\ entidad, \\ L - entidad, & si\ w_i\ es\ fin\ de\ entidad, \\ O, & no\ entidad, \\ U - entidad, & si\ w_i\ es\ entidad\ de \text{ único } elemento. \end{cases}$$

El efecto del uso de algún estilo de anotación en particular se da en dos aspectos. Primero, el número de clases que se manejará, ya que al diferenciar entre el inicio de una entidad, habrá que hacerlo para todas las entidades que contenga el corpus, de la misma forma con los demás prefijos (B,I,O,L,U). Segundo, en las métricas de evaluación, ya que para considerar que un clasificador ha etiquetado correctamente se toma en cuenta que toda la entidad esté correctamente reconocida, en caso que alguno de sus elementos no sea correcto, tal anotación es considerada como incorrecta; impactando directamente en la *Precision*, *Recall* y *F-measure* (las métricas de evaluación serán explicadas en la siguiente subsección).

En el Cuadro 2.2 se aprecian ejemplos de los estilos de anotación IO, BIO, BILOU, mostrando la forma en la que son utilizados. Cabe señalar que el estilo de etiquetado IO es equivalente a utilizar directamente el nombre de la entidad (como fue utilizado en el Cuadro 2.1) ya que en este estilo solo se diferencia entre un elemento dentro de una entidad o no.

2.2.2. Métricas de evaluación

Para realizar la evaluación de esta tarea, se hace uso de las métricas (Jurafsky y Martin, 2009) *Precision* (ver Ecuación 2.1), *Recall* (ver Ecuación 2.2) y *F-measure* (ver Ecuación 2.4). Para ello es necesario tomar en cuenta el Cuadro 2.3 (Manning y Schütze, 1999) en el que se resume el rendimiento del clasificador comparando las

Entidad	**IO**	**BIO**	**BILOU**
BCP	I-ORG	B-ORG	U-ORG
(	O	O	O
Banco	I-ORG	B-ORG	B-ORG
de	I-ORG	I-ORG	I-ORG
Crédito	I-ORG	I-ORG	I-ORG
del	I-ORG	I-ORG	I-ORG
Perú	I-ORG	I-ORG	L-ORG
)	O	O	O
,	O	O	O
La	I-ORG	B-ORG	B-ORG
Ibérica	I-ORG	I-ORG	L-ORG

Cuadro 2.2: Estilos de anotación

etiquetas que en realidad debieron ser asignadas (dos últimas columnas en el Cuadro 2.3) y las etiquetas que el clasificador NER ha asignado (dos últimas filas en el Cuadro 2.3).

	En realidad	
Clasificador NER	Correcto	Incorrecto
Seleccionado	tp	fp
No seleccionado	fn	tn

Cuadro 2.3: Matriz de contingencia

Los valores tp, fp, fn, tn hacen referencia a acumuladores del *resultado de clasificación* (sobre los ejemplos de evaluación) de verdadero positivo (*true positive*), falso positivo (*false positive*), falso negativo (*false negative*) y verdadero negativo (*true negative*).

El *resultado de clasificación* se obtiene de cada palabra del texto con respecto a las clases definidas, es decir que por cada palabra se le asignará como etiqueta una clase y será contabilizada dependiendo del Cuadro 2.3.

$$precision = \frac{tp}{tp + fp} \tag{2.1}$$

$$recall = \frac{tp}{tp + fn} \tag{2.2}$$

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{2.3}$$

$$F_1 = \frac{2PR}{P + R} \tag{2.4}$$

F-measure es calculado utilizando la Ecuación 2.3, donde β es usado para dar mayor importancia a *precision* o *recall* dependiendo de su valor, como se muestra en la Ecuación 2.5.

$$\beta = \begin{cases} recall, & si\ \beta\ > 1, \\ precision, & si\ \beta\ < 1, \\ balanceado, & \beta = 1. \end{cases} \tag{2.5}$$

Entonces, asignando balanceadamente la importancia a *precision* y *recall*, sea $\beta = 1$ obtenemos la Ecuación 2.4. Por consiguiente, de aquí en adelante se utilizará *F1* para referirse a *F-measure* con $\beta = 1$.

2.3. Clasificación de secuencias

Como se vio en la sección anterior, para poder realizar NER, es necesario definir un proceso de clasificación que permita predecir una serie de etiquetas para una oración dada. Esta tarea de clasificación, comúnmente denominada de secuencias emplea el aprendizaje supervisado de secuencias, que puede ser formulado como sigue. Sea $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{T}$ un conjunto de T ejemplos de entrenamiento. Cada ejemplo representa un par de secuencias $(\bar{x}_i, \bar{y}_i)$, donde $\bar{x}_i = \langle x_{i,1}, x_{i,2}, \ldots, x_{i,N_i} \rangle$ y $\bar{y}_i = \langle y_{i,1}, y_{i,2}, \ldots, y_{i,N_i} \rangle$ (Dietterich, 2002). Por ejemplo, en reconocimiento de entidades nombradas, (x_i, y_i) es un par que consiste de:

$\bar{x}_i = \langle Backus\ ,\ La\ Ibérica\ y\ BCP\ permanecen\ más\ de\ un\ siglo\ en\ el\ mercado\ local\ . \rangle$,
y
$\bar{y}_i = \langle org\ ne\ org\ org\ ne\ org\ ne\ ne\ ne\ amount\ amount\ ne\ ne\ ne\ ne\ ne \rangle$.

El objetivo es construir un clasificador h que correctamente pueda predecir una nueva secuencia de etiquetas $y = h(x)$ dada una oración de entrada x. El clasificador h puede ser representado por modelos gráficos probabilísticos para secuencias, como HMM y CRF (Klinger y Tomanek, 2007).

2.3.1. Hidden Markov Model

HMM es un enfoque de clasificación de secuencias generativo[2], que define una distribución de probabilidad conjunta $p(X, Y)$ (ver Ecuación 2.6), donde X y Y son variables aleatorias que representan, respectivamente, secuencias de observación y sus estados correspondientes (Wallach, 2004).

[2]Modelos generativos explícitamente intentan modelar una distribución de probabilidad conjunta $p(y, x)$ donde x son las observaciones y y los estados asociados a dichas observaciones.

$$p(X, Y) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_y|y_t) \tag{2.6}$$

En este modelo gráfico probabilístico se emplea la propiedad de Markov para considerar una cantidad k de *estados* de los cuales dependerá el estado actual, es decir, con esta propiedad se establece la independencia del estado actual con el resto de estados excepto el inmediato anterior (ver Figura 2.1).

En la Figura 2.1 se aprecia las dependencias en un HMM, donde cada observación x_i depende únicamente de su estado y_i y cada estado depende únicamente del estado anterior y_{i-1} (propiedad de Markov).

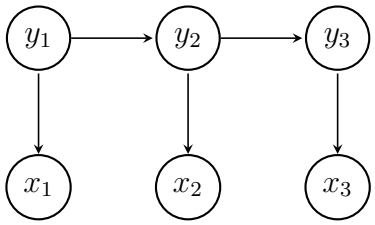


Figura 2.1: Modelo gráfico de HMM para etiquetado de secuencias.

Con fines de explicación, se considerará dos estados de dependencia en la propiedad de Markov que corresponde a un HMM de 2.º orden. La distribución de probabilidad conjunta representada por un modelo HMM se define en la Ecuación 2.7 mediante la probabilidad de transición denotada por q y la probabilidad de emisión denotada por e. La probabilidad de transición representa la probabilidad que los estados y_i, y_{i-1}, y_{i-2} aparezcan de forma conjunta, mientras que la probabilidad de emisión representa la probabilidad de la observación actual dado su estado.

La probabilidad de transición es definida en la Ecuación 2.8, donde $\sum_{i=1}^{3} \lambda_i = 1$. Mientras que, la probabilidad de emisión e es definida en la Ecuación 2.9, donde $\mathcal{F}(.)$ denota la frecuencia.

$$p(x_1, ..., x_n, y_1, ..., y_{n+1}) = \prod_{i=1}^{n+1} q(y_i|y_{i-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i|y_i) \tag{2.7}$$

$$q(y_i|y_{i-2}, y_{i-1}) = \lambda_1 p(y_{i-2}|y_i, y_{i-1}) + \lambda_2 p(y_{i-1}|y_i) + \lambda_3 p(y_i) \tag{2.8}$$

$$e(x_i|y_i) = \frac{\mathcal{F}(y_i, x_i)}{\mathcal{F}(y_i)} \tag{2.9}$$

En la Ecuación 2.8 se modela la probabilidad de la ocurrencia del n-grama[3] .

[3] Un n-grama es el contexto local de un *token*, en este caso de una palabra, entonces a partir de la palabra actual se observa una y dos palabras hacia atrás.

La probabilidad $p(y_{i-2}|y_i, y_{i-1})$ define un trigrama (ver Ecuación 2.10), la probabilidad $p(y_{i-1}|y_i)$ un bigrama (ver Ecuación 2.11) y $p(y_i)$ un unigrama (ver Ecuación 2.12).

$$p(y_{i-2}|y_i, y_{i-1}) = \frac{\mathcal{F}(y_{i-2}, y_{i-1}, y_i)}{\mathcal{F}(y_{i-2}, y_{i-1})} \tag{2.10}$$

$$p(y_{i-1}|y_i) = \frac{\mathcal{F}(y_{i-1}, y_i)}{\mathcal{F}(y_{i-1})} \tag{2.11}$$

$$p(y_i) = \frac{\mathcal{F}(y_i)}{\sum_{i=1} \mathcal{F}(y_i)} \tag{2.12}$$

En el modelo definido se observan las etiquetas (probabilidad de transición) y palabras (probabilidad de emisión), sin tomar en cuenta alguna otra característica en el texto. Brants (Brants, 2000) propone ligeras modificaciones a la definición de las probabilidades para considerar una característica adicional c_i, para cada elemento x_i de la secuencia de extrada $x_1, ..., x_n$ (si la palabra actual esta en mayúscula o no) de la siguiente forma:

$$p(x_1, ..., x_n, y_1, ..., y_{n+1}) = \prod_{i=1}^{n+1} q(y_i, c_i|y_{i-2}, c_{i-2}, y_{i-1}, c_{i-1}) \prod_{i=1}^{n} e(x_i|y_i, c_i) \tag{2.13}$$

Con estas modificaciones, la probabilidad de transición de la Ecuación 2.8 ahora es representada por la Ecuación 2.14, donde se añaden dos sumandos con λ_4 (para trigrama con la característica, desarrollada en la Ecuación 2.15) y λ_5 (para bigrama con la característica, desarrollada en la Ecuación 2.16) para modelar la probabilidad de la etiqueta y_i y que la palabra esté en mayúscula o no, denotada por c_i.

$$\begin{aligned} q(y_i, c_i|y_{i-2}, c_{i-2}, y_{i-1}, c_{i-1}) =& \lambda_1 p(y_{i-2}|y_i, y_{i-1}) + \lambda_2 p(y_{i-1}|y_i) + \lambda_3 p(y_i) + \\ & \lambda_4 p(y_{i-2}, c_{i-2}|y_i, c_i, y_{i-1}, c_{i-1}) + \\ & \lambda_5 p(y_{i-1}, c_{i-1}|y_i, c_i) \end{aligned} \tag{2.14}$$

$$p(y_{i-2}, c_{i-2}|y_i, c_i, y_{i-1}, c_{i-1}) = \frac{\mathcal{F}(y_{i-2}, c_{i-2}, y_{i-1}, c_{i-1}, y_i, c_i)}{\mathcal{F}(y_{i-2}, c_{i-2}, y_{i-1}, c_{i-1})} \tag{2.15}$$

$$p(y_{i-1}, c_{i-1}|y_i, c_i) = \frac{\mathcal{F}(y_{i-1}, c_{i-1}, y_i, c_i)}{\mathcal{F}(y_{i-i}, c_{i-1})} \tag{2.16}$$

Por otro lado, la probabilidad de emisión de la Ecuación 2.9 es modificada por la Ecuación 2.17, donde γ_1 y γ_2 suman 1.

$$e(x_i|y_i, c_i) = \gamma_1 \frac{\mathcal{F}(y_i, x_i)}{\mathcal{F}(y_i)} + \gamma_2 \frac{\mathcal{F}(y_i, c_i, x_i)}{\mathcal{F}(y_i, c_i)} \tag{2.17}$$

2.3.2. Conditional Random Fields

Para definir *Conditional Random Fields* se utilizará las Redes de Markov (Koller y Friedman, 2009). Una Red de Markov es un modelo gráfico probabilístico con una estructura dada por un grafo no dirigido. Los nodos en el grafo representan las variables y las asociaciones corresponden a la directa interacción probabilística entre las variables.

Su distribución de probabilidad se define por la Distribución de Gibbs. Una distribución P_Φ es una Distribución de Gibbs parametrizada por un conjunto de factores $\Phi = \{\phi_1(D_1), ..., \phi_K(D_K)\}$, definida como:

$$P_\Phi(X_1, ..., X_n) = \frac{1}{Z}\tilde{P}_\Phi(X_1, ..., X_n) \tag{2.18}$$

Donde,

$$\tilde{P}_\Phi(X_1, ..., X_n) = \phi_1(D_1) \times \phi_2(D_2) \times ... \times \phi_m(D_m) \tag{2.19}$$

es una medida sin normalizar y

$$Z = \sum_{X_1,...,X_n} \tilde{P}_\Phi(X_1, ..., X_n) \tag{2.20}$$

es una constante llamada función de partición, cada $\phi_i(D_i)$ representa el factor de los nodos asociados.

Luego, se dice que una distribución P_Φ con $\Phi = \{\phi_1(D_1), ..., \phi_K(D_K)\}$ factoriza una Red de Markov $\mathcal{H}$ si cada D_K $(k = 1, ..., K)$ es un subgrafo completo de $\mathcal{H}$

CRF es un modelo gráfico probabilístico representado por un grafo no dirigido (ver Figura 2.2), parametrizado de la misma forma que una Red de Markov (Koller y Friedman, 2009), como un conjunto de factores $\phi_1(D_1), ..., \phi_m(D_m)$, pero ahora, la distribución de probabilidades resolverá una distribución condicional $P(\boldsymbol{Y}|\boldsymbol{X})$, donde $\boldsymbol{Y}$ y $\boldsymbol{X}$ representan a un conjunto de variables aleatorias respectivamente. La distribución de probabilidad de CRF se define de la siguiente forma:

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})}\tilde{P}(\boldsymbol{Y}, \boldsymbol{X}) \tag{2.21}$$

$$\tilde{P}(\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{m} \phi_i(D_i) \tag{2.22}$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \tilde{P}(\boldsymbol{Y}, \boldsymbol{X}) \tag{2.23}$$

Más concretamente, este modelo discriminativo define la distribución de probabilidad condicional $P(\boldsymbol{Y}|\boldsymbol{X})$ de una salida $\boldsymbol{Y}$ dada la entrada $\boldsymbol{X}$ también llamada *observación* (Lafferty et al., 2001). En el caso de la tarea NER, $\boldsymbol{X}$ son las características de la secuencia y $\boldsymbol{Y}$ la categoría de dicha secuencia (por ejemplo B-PER, I-PER, B-ORG), entonces CRF modela la probabilidad condicional de la clase dadas las características observadas.

Sea $D_i = \{Y_i, X_i\}$, entonces,

$$\phi_i(D_i) = exp(\sum_{k=1}^{n} \lambda_k f_k(y_{j-1}, y_j, \boldsymbol{X_i}, j)) \tag{2.24}$$

Como las redes de Markov, la distribución de probabilidad (Klinger y Tomanek, 2007) de CRF (ver Ecuación 2.21) está formada por factores ($\phi_i(D_i)$) y por una función de normalización ($Z(\boldsymbol{X})$), para garantizar que se tenga una distribución de probabilidades legal.

El modelo gráfico se muestra en la Figura 2.2, formado por la secuencia a etiquetar (denotado por s_i) y la característica de la secuencia (denotada por c_i). En este modelo gráfico se aprecia las dependencias de las variables. Vemos que s_2 depende de las etiquetas (clases) s_1 y de s_3, asimismo de la observación (características de la secuencia) c_2, de forma similar ocurre con s_1 y s_3

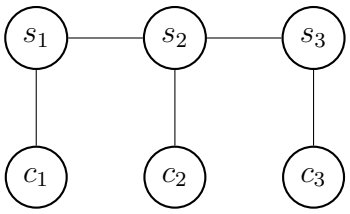


Figura 2.2: Modelo gráfico de la cadena lineal CRF para etiquetado de secuencias.

Las características de las secuencias son rasgos particulares que se observan y están dadas por los factores ($\phi_i(D_i)$). En la Ecuación 2.24, el término $f_k(y_{j-1}, y_j, X_i, j)$ hace referencia a las características de la observación actual X_i en la posición j con etiqueta actual y_j y etiqueta de la observación previa y_{j-1}. Mediante f_k se define la característica que se espera cumplir en la secuencia observada. El término λ_k representa el peso de la característica f_k, aprendida durante el entrenamiento del modelo sobre el conjunto de datos.

En CRF se observa un conjunto de características de la secuencia a etiquetar, para definir su distribución de probabilidad condicional (como se aprecia en la Ecuación 2.21). Las características pueden ser dadas tomando en cuenta las observaciones

(palabras) o dentro de la misma observación (caracteres en la palabra) (Lafferty et al., 2001). Debido a la capacidad de modelar la distribución de probabilidad condicional de la clase dadas las características observadas (modelo discriminativo), CRF alcanza mejor rendimiento que HMM.

2.4. Características en secuencias de textos

El texto escrito se presenta de forma natural bajo ciertas características, mayormente establecidas por la gramática de un idioma en particular o por el dominio (por ejemplo en textos farmacológicos). Las características que provienen de rasgos particulares de un idioma pueden ser consideradas como *conocimiento experto* dado que reflejan el uso de una lengua.

El *conocimiento experto* puede llevar a formular características que serán llamadas ***características supervisadas*** por su naturaleza. Por ejemplo, los nombres propios en el idioma español son escritos con la primera letra en mayúscula. Otro claro ejemplo lo encontramos en NER, en el uso de un léxico (comúnmente llamado *"gazetteer"*) que contiene conjuntos de palabras asociados a una entidad en particular, más concretamente, un léxico de nombres de personas (Barceló et al., 2009).

Sin embargo, también existen rasgos más complejos de identificar, muy dependientes de cada idioma y de la semántica del contexto de las palabras que pueden ser conseguidos con un *profundo conocimiento experto*. Este conjunto de características son tratadas con técnicas de aprendizaje no supervisado las que serán llamadas como ***características no supervisadas***.

Las características no supervisadas exploradas están basadas en *word representations* (*clustering* de palabras y *word embeddings*) y colocaciones.

2.4.1. *Word representations*

Una *word representation* es una forma en la que una palabra es representada, mediante cálculos matemáticos por lo que es conocida como un objeto matemático[4] asociado con cada palabra, frecuentemente un vector de valores continuos . Cada valor corresponde a una característica e incluso tiene una interpretación semántica o gramática, es así que es posible utilizarlo como característica de la palabra en un texto (Turian et al., 2010).

Las *word representations* que han alcanzado mejores resultados han sido las que se basan en clustering de palabras y en *word embeddings* (Turian et al., 2010). A continuación se explica la forma en que cada *word representation* es calculada.

[4]Tal como números, funciones, espacio de vectores.

- ***Brown clustering*** El *Brown clustering* es un clustering jerárquico de palabras que toma una secuencia de palabras $w_1, \ldots, w_n$ como entrada y retorna un árbol binario como resultado. Las hojas del árbol binario son las palabras. Este método de *clustering* está basado en los modelos de lenguaje de bigrama calculados mediante la fórmula en la Ecuación 2.25 (Brown et al., 1992; Liang, 2005).

$$p(w_1, w_2, ..., w_T) = \prod_{i=1}^{T} p(w_i|C(w_i))p(C(w_i)|C(w_{i-1})) \tag{2.25}$$

  Donde $C(w_i)$ denota el *cluster* de la palabra w_i. A continuación se presenta un texto de ejemplo: "La multinacional española Telefónica ha impuesto un récord mundial al poner en servicio tres millones de nuevas líneas en el estado brasileño de Sao Paulo desde que asumió el control de la operadora Telesp hace 20 meses, anunció hoy el presidente de Telefónica do Brasil, Fernando Xavier Ferreira." y su árbol binario se aprecia en la Figura 2.3. En este breve ejemplo se observa que palabras que ocurren en contextos similares son asignadas al mismo *cluster*, por mencionar dos casos, en el *cluster* 0101 (hoja del árbol) se encuentran las palabras 'Fernando' y 'Pereira' que son mencionadas en el nombre de una persona, luego en el cluster 01101 están las palabras 'que' y 'en' las cuales ocurren en un texto para relacionar ideas o palabras.

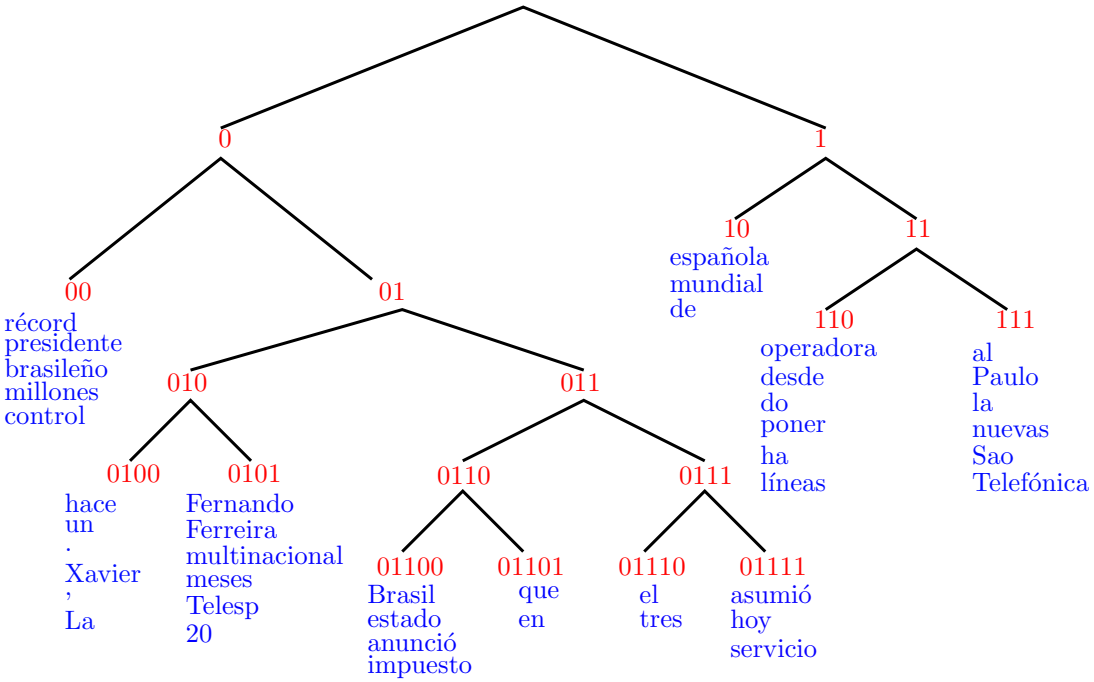


Figura 2.3: Árbol binario de *Brown clusters*

- ***Word embeddings*** Los *word embeddings* son representaciones de palabras en vectores, definidos como densos, continuos y de pocas dimensiones (Turian et al., 2010; Guo et al., 2014). Estos vectores pueden ser aprendidos de textos sin etiquetar a través de modelos de predicción del contexto o métodos espectrales de forma no supervisada (Guo et al., 2014).

Cada dimensión del vector representa una característica latente de la palabra preservando sus regularidades lingüísticas. De esta forma, se espera que palabras similares estén distribuidas próximas en el espacio (Guo et al., 2014).

Considerando el modelo *Skip-gram* (Mikolov et al., 2013b,a) (modelo de predicción de contexto más eficiente (Mikolov et al., 2013b; Guo et al., 2014)), se toma como entrada la palabra actual w y se predice la distribución de probabilidad del contexto de palabras (observando un número de palabras previo y posterior de la palabra actual) (Guo et al., 2014).

El modelo *Skip-gram* está dado por una red neuronal con tres capas, como se aprecia en la Figura 2.4. En la capa de entrada (*input layer*) se hace uso de una representación *sparse* de la palabra, luego a través de pesos se proyecta la palabra (*projection layer*) en la siguiente capa para predecir cada palabra que pertenece a su contexto.

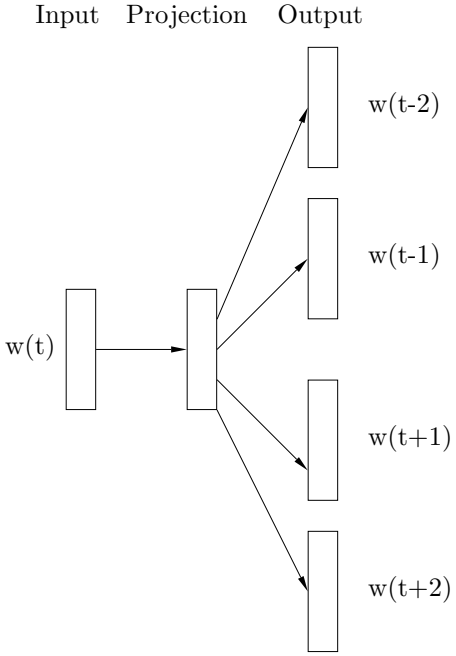


Figura 2.4: Arquitectura del modelo *Skip-gram* (Mikolov et al., 2013b,a)

Dada una secuencia de entrenamiento w_1, w_2, w_3,..., w_T (cada w_i representa una palabra en un texto), el objetivo del modelo *Skip-gram* es maximizar:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} log\ p(w_{t+j}|w_t) \tag{2.26}$$

Donde c es el tamaño del contexto de entrenamiento (el cual está en función de la palabra central w_t) y T es el tamaño del vocabulario. La formulación básica del modelo *Skip-gram* define $p(w_{t+j}|w_t)$ utilizando la función softmax:

$$p(w_s|w_e) = \frac{exp({v'}_{w_s}^T v_{w_e})}{\sum_{w=1}^{W} exp({v'}_{w}^T v_{w_e})} \tag{2.27}$$

Entonces, como resultado del cálculo de las *word embeddings* se tiene un vector que representa a cada palabra en el vocabulario del texto con que ha sido entrenado. Por otro lado, el uso directo de los *word embeddings* como características ha mostrado no alcanzar los mejores resultados en este tipo de enfoques (Turian et al., 2010), por ello que se ha estudiado la forma de aplicar estas *características de las palabras* para alcanzar mejores resultados (Guo et al., 2014). En la literatura se ha estudiado el uso de los enfoques: Binarización de *embeddings*, *clustering* de *embeddings* y prototipos distribucionales.

- ***Embeddings* Binarizados** La idea detrás de este método es "reducir" los vectores continuos de las palabras $\vec{w}$ (de los *word embeddings*), a vectores discretos $bin(\vec{w})$, que sin embargo conserven los rasgos más resaltantes de los *embeddings*. Con este fin, es necesario calcular dos umbrales por dimensión (superior e inferior) a través de todo el vocabulario. Por cada dimensión (componente del vector) i es calculado el *promedio* de los valores positivos (C_{i+}, el umbral superior) y valores negativos (C_{i-}, el inferior). Luego, la siguiente función es utilizada sobre cada componente C_{ij} del vector $\vec{w}_j$:

$$\phi(C_{ij}) = \begin{cases} U_+, & if C_{ij} \geq promedio(C_{i+}), \\ B_-, & if C_{ij} \leq promedio(C_{i-}), \\ 0, \end{cases} \tag{2.28}$$

  Al aplicar la Ecuación 2.28 se consigue un vector con valores entre $\{U+, B-, 0\}$, pero al momento de utilizarlos como características sólo se considera los valores U+ y B-, por ello el nombre de *binarización de embeddings*.

- ***Clustering* de *embeddings*** Un método de *clustering* de *embeddings* basado en *k-means* fue propuesto por Yu et al. (Yu et al., 2013). Experimentos realizados han demostrado que diferentes números para *k's* contienen información de diferente granularidad.

- **Prototipos Distribucionales**
  Este enfoque esta basado en la idea de que cada clase (B-PER, I-PER, B-ORG, I-ORG, entre otros) tiene un conjunto de palabras que es más probable que pertenezcan a ella (por ejemplo Maria, Jose son palabras que tienen una mayor probabilidad de ser clasificadas como la entidad B-PER). De esta manera, resulta útil identificar un grupo de palabras que representen a cada clase (que son los *prototipos*) y seleccionarlas como rasgos o características, pero solamente aquellos más parecidos (usando la similaridad del coseno).
  Para calcular los prototipos Guo et al. (Guo et al., 2014) define dos pasos necesarios:

  1. Generar los prototipos por cada clase a partir de un corpus de entrenamiento anotado. En este paso se hace uso de *Normalized Pointwise Mutual Information* (NPMI) (Bouma, 2009). El tipo de relación *palabra-entidad* puede ser modelado como una forma de colocación. NPMI es una versión suavizada de la medida *Mutual Information* tipicamente usada para detectar asociaciones (Yang y Pedersen, 1997) y colocaciones (Liang, 2005). Dado un corpus anotado de entrenamiento, NPMI

se calcula a partir de las etiquetas l y las palabras w utilizando las siguientes dos fórmulas:

$$\lambda_n(l,w) = \frac{\lambda(l,w)}{-\ln p(l,w)}, \quad \lambda(l,w) \ = \ln \frac{p(l,w)}{p(l)p(w)}.$$

2. Mapear los prototipos a palabras (en los *word embeddings*). En este paso, dado un grupo de prototipos para cada clase, se busca cuáles de esos prototipos son más *similares* a cada palabra en los embeddings. Se utiliza la similaridad del coseno y aquellos prototipos por encima de un umbral (usualmente 0,5), son elegidos como características prototipo de la palabra.

2.4.2. Colocaciones

Se define como colocación, cuando dos o más elementos léxicos ocurren frecuentemente en un texto o en un corpus, sin importar si forman o no un patrón sintáctico (Poulsen, 2005). El cálculo de las colocaciones en datos no etiquetados ha sido realizado por conteos de bigramas e información mutua, dada por la Ecuación 2.29 (Liang, 2005).

$$MI(w_i, w_{i+1}) = log \frac{p(w_i w_{i+1})}{p(w_i)p(w_{i+1})} \tag{2.29}$$

2.5. Consideraciones Finales

En este capítulo se han presentado conceptos necesarios que más adelante serán abordados. Se definió NLP y NER, y también se explicó la forma en que se lleva a cabo la clasificación de secuencias y las características en las secuencias de texto. En el siguiente capítulo se presentan los trabajos relacionados.

Capítulo 3

Trabajos Relacionados

En esta sección se presentan los trabajos relacionados con respecto a dos tópicos. Primero, sobre el reconocimiento de entidades nombradas en el idioma español. Luego, sobre el uso de características no supervisadas en el modelo gráfico probabilístico CRF para el idioma inglés (ya que este tipo de estudios aún no se han hecho sobre el español).

3.1. NER en el idioma español

El estudio de entidades nombradas para el español se ha dado principalmente bajo dos modalidades. En primer lugar, reconociendo varios tipos de entidades en el texto a la vez (como *Person*, *Location*, *Organization*). En la segunda modalidad, se estudia de forma aislada alguno de los tipos de entidades.

Con respecto a la ***primera modalidad*** de NER en el idioma español, la tarea ha sido estudiada empleando un conjunto de características (definidas de forma manual) y un conjunto de palabras asociadas a cada entidad (léxico comúnmente denominado "gazetteers") como parte de un modelo de aprendizaje Adaboost (Carreras et al., 2002) y CRF (Finkel et al., 2005) sobre el corpus CoNLL 2002 en español.

A pesar de que Stanford CRF NER (Finkel et al., 2005) no fue creado para un idioma en particular, el conjunto de características (supervisadas y/o no supervisadas) para el idioma español aún no ha sido estudiado. Mientras que para el idioma Inglés (Finkel et al., 2005), Alemán (Faruqui y Padó, 2010) y Chino (Che et al., 2013) las características si han sido estudiadas.

Luego, la tarea ha sido abordada con otro grupo de modelos de aprendizaje basados en aprendizaje profundo (*Deep Learning*) sobre el corpus CoNLL 2002 en español empleando principalmente características (aprendidas sin supervisión) a dos niveles de granularidad:

1. Palabra: a través de *word embeddings*.

2. Carácter: a través de *character embeddings*.

Las principales arquitecturas en aprendizaje profundo que han sido utilizadas en la tarea están dadas por *Convolutional Neural Network* (CNN) (Redes Neuronales Convolutivas) y *Recurrent Neural Network* (RNN) (Redes Neuronales Recurrentes). CNN está compuesto por módulos de capas convolutivas y de agrupamiento (*pooling layer*). Estos módulos son apilados frecuentemente y las capas convolutivas comparten pesos, la capa de agrupamiento reduce los datos. CNN aplicado a textos (Deng y Yu, 2014) produce características locales alrededor de cada palabra (o de cada *token*) de la oración gracias a las capas convolutivas que son combinadas en un vector de características global (Collobert et al., 2011). RNNs son una familia de redes neuronales que operan en datos secuenciales tomando como entrada una secuencia de vectores y retornando otra secuencia que representa alguna información sobre la secuencia (Lample et al., 2016).

Los estudios realizados en NER para el idioma español han empleado CNN con *word* y *character embeddings* (dos Santos y Guimarães, 2015), *RNN* con *word* y *character embeddings* (Lample et al., 2016; Yang et al., 2016), *CNN* empleando *word* y *character embeddings* conjuntamente con *RNNs* (Murthy y Bhattacharyya, 2016) y una RNN basada en una codificación a nivel de bytes (Gillick et al., 2015).

Por un lado, tenemos que CRF es el modelo de aprendizaje de secuencias tradicional más exitoso (Finkel et al., 2005; Glavaš et al., 2012). De otro lado, la arquitectura con mayor poder predictivo en el aprendizaje profundo es RNN. Luego, recientes estudios han combinado estos dos modelos de aprendizaje de secuencias (Lample et al., 2016; Yang et al., 2016) alcanzando el estado del arte actual para la tarea.

Los modelos de aprendizaje hasta este punto se han entrenado en el conjunto de entrenamiento de CoNLL-2002 y evaluados con su conjunto de prueba. Sin embargo, otro reciente estudio ha sido llevado a cabo para construir un corpus anotado de mayor tamaño para el reconocimiento de entidades nombradas multi-idioma y evaluado en el conjunto de prueba de CoNLL-2002 (para evaluar el rendimiento en el idioma español) con un modelo de aprendizaje basado en redes neuronales (Al-Rfou et al., 2015) (sin emplear aprendizaje profundo) dando como resultado un conjunto de entidades nombradas por idioma (para el idioma español es conocido como *wiki-3*). En el Cuadro 3.1 se muestra el estado del arte para el idioma español con los modelos respectivamente estudiados y evaluados en el corpus CoNLL-2002 en español.

Un fenómeno que ocurre en entidades de ciertos dominios son las entidades anidadas (entidades contenidas en otras). Finkel y Manning (Finkel y Manning, 2009) trataron este problema para el idioma español utilizando un analizador de constituyentes discriminativo evaluado sobre AnCora en los idiomas español y catalán. En este estudio se expandió la influencia, no solo a las palabras alrededor de las entidades sino a las entidades contenidas.

Con respecto a la ***segunda modalidad*** de NER para el idioma español, la cual está enfocada en propuestas que profundizan el reconocimiento de entidades nombradas

Modelo	*F1*
Polyglot (Al-Rfou et al., 2015)	62.37 %
Adaboost (Carreras et al., 2002)*	79.28 %
Adaboost (Carreras et al., 2002)	81.39 %
CNN para palabras y caracteres (dos Santos y Guimarães, 2015)	82.21 %
CNN + RNN (Murthy y Bhattacharyya, 2016)	82.59 %
RNN (Gillick et al., 2015)	82.95 %
RNN + CRF (Lample et al., 2016)	85.75 %
Gated RNN + CRF (Yang et al., 2016)	85.77 %

Cuadro 3.1: Estado del arte en CoNLL-2002 español

*sin tomar en cuenta gazetteers

a una sola entidad en un corpus, se estudió de forma sesgada las entidades nombradas (solamente el tipo *"Person"*). Mediante la definición de una gramática para reconocer nombres hispanos sobre un corpus de nombres mexicanos, bajo la premisa que los nombres hispanos en su mayor parte están formados por dos nombres y dos apellidos (materno y paterno), asimismo haciendo uso de *gazetteers* (Barceló et al., 2009). También mediante el uso de conocimiento heterogéneo para unir o separar grupos de palabras con mayúsculas para reconocer nombres de personas anidados en español, se recopiló artículos periodísticos de periódicos mexicanos y fueron utilizados en la evaluación (Galicia-Haro et al., 2004).

3.2. Características no supervisadas

El uso de características no supervisadas (en el idioma inglés) para modelar el reconocimiento de entidades nombradas ha sido propuesto previamente sobre una variedad de técnicas de aprendizaje máquina, tales como HMM (Miller et al., 2004), *Active Learning* (Ratinov y Roth, 2009), CRF (Liang, 2005; Finkel et al., 2005; Turian et al., 2010; Passos et al., 2014; Guo et al., 2014).

En particular, sobre el modelo gráfico probabilístico CRF hay estudios más recientes (Turian et al., 2010; Passos et al., 2014; Guo et al., 2014). Estos estudios exploraron el uso de *word representations* como características en un modelo CRF mediante: clustering de palabras (Liang, 2005) y *word embeddings* (Turian et al., 2010; Passos et al., 2014; Guo et al., 2014). Turian et al. (Turian et al., 2010) muestran que utilizar *clustering* de palabras permite caracterizar el patrón de etiquetado de mejor manera que *word embeddings*.

Analizando a mayor detalle los resultados alcanzados por Turian et al. (Turian et al., 2010), la forma en la que se modela la probabilidad de pertenencia entre un *token*[1] y un *cluster* está basada en bigramas, mientras que con *word embeddings* se modela la probabilidad de un *token* en un contexto lo cual hace que se represente de

[1]Elemento en el vocabulario del texto.

una manera más adecuada cada *token*. Este hecho debería ser reflejado en los resultados alcanzados por Turian et al. (Turian et al., 2010), sin embargo Guo et al. (Guo et al., 2014) y Passos et al. (Passos et al., 2014) exploran nuevamente estas características mostrando mejores maneras de utilizar los *word representations*.

Es así que Guo et al. (Guo et al., 2014) presentaron formas en las que se puede utilizar *word embeddings* como características en un modelo CRF logrando demostrar que para el idioma inglés esta *word representation* es mejor que el *cluster* de palabras. Otra forma de alcanzar resultados competitivos con el estado del arte con el modelo CRF es a través de un grado de granularidad de *embeddings* superior a las mencionadas hasta ahora, a nivel de frases (Passos et al., 2014).

Finalmente, las colocaciones de palabras han sido utilizadas para la segmentación de palabras en el idioma chino (Liang, 2005) calculadas de forma no supervisada a través de información mutua.

Word representations y colocaciones como parte de las características en un modelo de aprendizaje CRF aún no han sido exploradas para el idioma español y es una parte importante de la propuesta de este trabajo.

3.3. Consideraciones Finales

En esta sección se ha mostrado el estado del arte para NER en el idioma español y cómo se ha llevado a cabo la adición de características en el modelo gráfico probabilístico CRF. Cabe señalar que este tipo de estudios no han sido llevados a cabo para el español y dado que el estudio de esta tarea es relativa al idioma, acentúa aún mas la importancia de este trabajo. En el siguiente capítulo se presenta la propuesta mediante el uso de características mínimamente supervisadas y características no supervisadas.

Capítulo 4

Propuesta

En este capítulo se presenta la propuesta de la tesis, asimismo, las características supervisadas y no supervisadas empleadas en el modelo gráfico probabilístico CRF y la forma en la que las características se configuran en él.

4.1. Modelo propuesto

Para reconocer entidades nombradas en el idioma español se plantea utilizar el modelo gráfico probabilístico CRF como clasificador de entidades haciendo uso de características supervisadas y no supervisadas (ver Figura 4.1), con énfasis en estas últimas.

Recapitulando, con CRF se modela la distribución de probabilidad condicional de una clase dado un conjunto de características. Este modelo discriminativo permite aprender secuencias mediante el conocimiento que se le proporciona (las características de la secuencia) y, asimismo, es el clasificador de secuencias con mejor rendimiento (Lafferty et al., 2001; Finkel et al., 2005; Glavaš et al., 2012; Bingel y Haider, 2014; Lample et al., 2016; Yang et al., 2016) ya que modela el conocimiento experto mediante las relaciones de dependencia entre las características de una secuencia dada y su etiqueta (entidad).

El conjunto de *características supervisadas* (definidas en la Sección 4.2) ha sido definido con el objetivo de representar rasgos de las palabras, por ejemplo en relación al tamaño del *token*, caracteres que componen el *token*. Por otro lado las *características no supervisadas* (definidas en la Sección 4.3) tienen por objetivo reflejar las relaciones entre palabras, por ejemplo asignando el mismo *cluster* a los verbos, encontrando la relación entre dos *tokens* por similitud. Asimismo, las características aprendidas de forma no supervisada están dadas por *word representations* (*clustering* de palabras, *word embeddings*) y de un recurso lingüístico (colocaciones).

Adicionalmente a las características ya mencionadas, se hace uso de un conjunto de datos en el idioma inglés, así esta propuesta utiliza características multilenguaje . Se

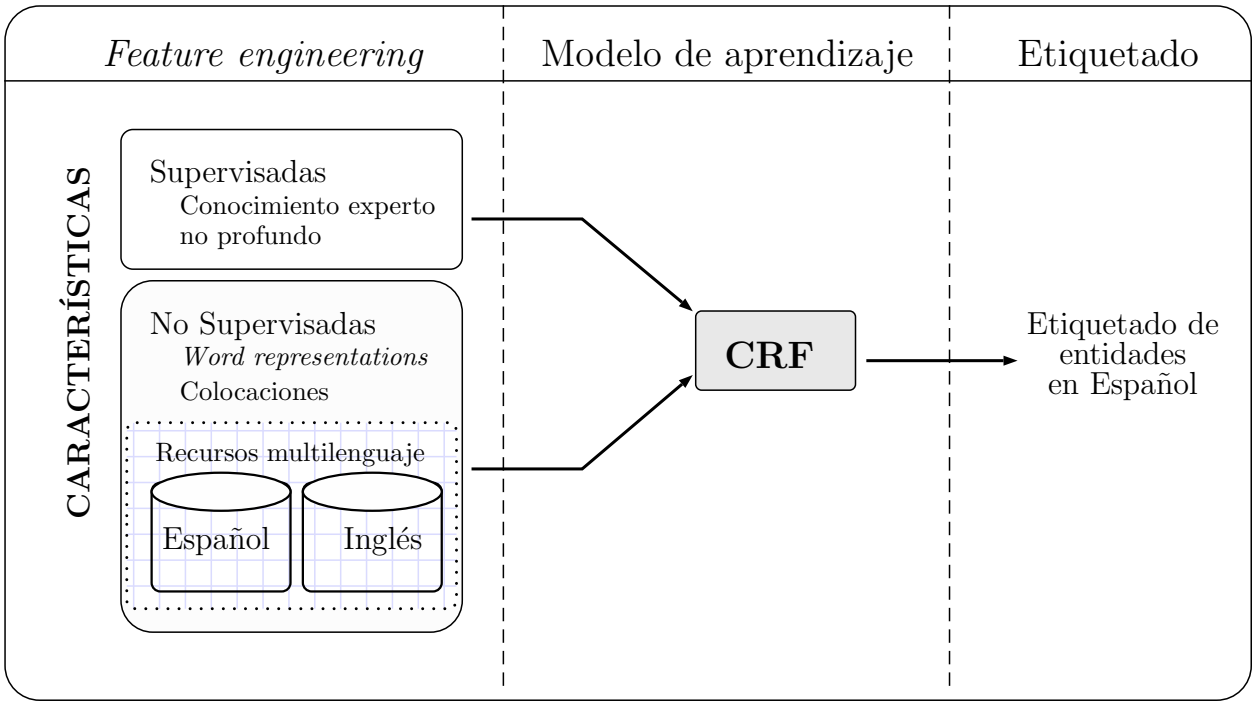


Figura 4.1: Esquema de la propuesta para NER en español

hace uso de este recurso ya que las entidades usualmente son nombradas de la misma forma en estos dos idiomas, por lo que al añadir este conjunto de datos, se expande la cobertura del vocabulario y por ende del contexto en que ocurren las palabras.

A continuación se especifica las características que han sido utilizadas en este trabajo, y en el caso de las características no supervisadas se explica cuál es la característica que finalmente se usa.

4.2. Características supervisadas

Se definen un conjunto de características supervisadas como características iniciales (de línea base) sobre una ventana de ± *2 tokens*. El conjunto de características tomadas para cada *token* se detallan a continuación:

- La palabra en sí.
- La palabra en minúscula.
- Etiqueta *part-of-speech*.
- Patrón del uso de mayúsculas (por ejemplo, de la palabra "Twitter" su patrón es ULLLLLL) y tipo de carácter en la palabra (por ejemplo, todo en mayúscula, todos son dígitos, todos son símbolos).

- Información del tipo de carácter: si está en mayúscula o minúscula, si es dígito o símbolo, o si la palabra tiene el primer carácter en mayúscula o todos los carácteres, o si todos los carácteres son letras o dígitos.
- Prefijos y sufijos: cuatro primeras o últimas letras respectivamente.
- Tamaño de dígito: si el token actual tiene tamaño dos o cuatro.
- Combinación con dígito: cual combinación entre el dígito y token actual está presente (alfanumérico, barra, coma, punto).
- Si el token actual contiene una letra en mayúscula y un punto, o si contiene alguna letra en mayúscula, minúscula, dígito, caracter alfanumérico, símbolo.
- Indicador si la primera letra está en mayúscula, si todos los caracteres están en mayúsculas, si todos los caracteres estan en minúsculas, si todos son dígitos, si todos los caracteres no son alfanuméricos.

En NER es una práctica común emplear *gazetteers* como parte de las características de línea base (supervisadas), sin embargo en este trabajo no se considera ese tipo de recursos con el fin de potenciar la tarea de reconocimiento de entidades con características no supervisadas.

4.3. Características no supervisadas

Las características no supervisadas empleadas son *clustering* de palabras (*Brown clustering*), *word embeddings* y colocaciones. Asimismo, se emplean conjuntos de datos en el idioma español e inglés.

Para utilizar *word embeddings* se hace uso de los enfoques *embeddings* binarizados, *clustering* de *embeddings* y prototipos distribucionales (definidos en la Subsección 2.4.1).

- **Brown clustering**

  En el Cuadro 4.1 se muestran algunos ejemplos de *Brown clusters*. La columna *Brown clusters* es utilizada como característica de la correspondiente palabra.

- ***Embeddings* Binarizados**

  En el Cuadro 4.2 se muestra una vista reducida del vector de la palabra "equipo". En la primera columna se muestra cada dimensión de "equipo" y la segunda es su valor correspondiente. La tercera columna muestra el valor binarizado por cada dimensión, el cual forma el vector binarizado.

Brown Clusters	Palabra
011100010	Française
011100010	Hamburg
0111100011010	latino
0111100011010	conservador
0111111001111	malogran
0111111001111	paralizaban
011101001010	Facebook
011101001010	Twitter
011101001010	Internet

Cuadro 4.1: Ejemplos de *Brown clusters*.

Dimension	Valor	Binarizado
1	-0.008255	0
2	0.145529	U+
3	0.010853	0
⋮	⋮	⋮
298	0.050766	U+
299	-0.066613	B-
300	0.073499	U+

Cuadro 4.2: *Embeddings* binarizados para la palabra "equipo".

- **Clustering de *embeddings***

  Con el fin de modelar diferentes niveles de granularidad se calculan diferentes tamaños de *clusters* (Guo et al., 2014). Como características para cada palabra w se le asigna cada nivel de granularidad. En el Cuadro 4.3 se muestra los *clusters* de los embeddings calculados para la palabra "Maria". La primera columna denota el nivel de granularidad. La segunda columna denota el *cluster* asignado a "Maria" en cada nivel de granularidad.

Granularidad	k
500	31
1000	978
1500	1317

Cuadro 4.3: *Clustering* de *embeddings* para la palabra "Maria".

- **Prototipos Distribucionales**

  Por cada clase en un corpus (anotado con las entidades nombradas) se calculan los k primeros prototipos (con respecto a NPMI). Para cada corpus se ha calculado el número k de prototipos evaluados en su respectivo conjunto de desarrollo (*development set*).

  Con el fin de mostrar algunos ejemplos, en el Cuadro 4.4 se muestra los primeros cuatro prototipos por cada clase en el corpus CoNLL 2002 en español (del con-

junto de entrenamiento). Estos prototipos son instancias de cada clase (incluso de la clase correspondiente a las palabras que no son entidades (O)), por tanto, están compuestos por entidades o partes de entidades, tal es el caso de la entidad *Buenos Aires* (LOC), la palabra *Aires* es un prototipo de la clase I-LOC.

Clase	**Prototipos**
B-ORG	EFE, Gobierno, PP, Ayuntamiento
I-ORG	Nacional, Europea, Unidos, Civil
I-MISC	Campeones, Ambiente, Ciudadana, Profesional
B-MISC	Liga, Copa, Juegos, Internet
B-LOC	Madrid, Barcelona, Badajoz, Santander
I-LOC	Janeiro, York, Denis, Aires
B-PER	Francisco, Juan, Fernando, Manuel
I-PER	Alvarez, Lozano, Bosque, Ibarra
O	que, el, en, y

Cuadro 4.4: Prototipos calculados.

- **Colocaciones**

  Las colocaciones calculadas fueron asociadas a cada palabra en el corpus de tal forma que se han convertido en características. El Cuadro 4.5 muestra ejemplos para las palabras “Estados” y “General”.

Palabra	**Colocaciones**	
Estados	los	miembros
	Miembros	Unidos
General	Asamblea	Secretario

Cuadro 4.5: Colocaciones calculadas de las palabras: “Estados” y “General”.

4.4. Consideraciones Finales

En este capítulo se ha presentado la propuesta de la tesis, mostrando las características empleadas (supervisadas y no supervisadas) en el modelo gráfico probabilístico CRF y explicando la forma en la que las características serán utilizadas. Cabe resaltar que en este trabajo, se utiliza características supervisadas sin recursos externos, y se utilizan características no supervisadas con un enfoque multilenguaje para NER en español.

En la siguiente sección se presentan los conjuntos de datos, forma de evaluación y resultados sobre el modelo propuesto.

Capítulo 5

Resultados

En este capítulo se presenta la metodología de evaluación, explicando los conjuntos de datos (con/sin anotaciones) y detalles de implementación a tomar en cuenta. Finalmente, los resultados obtenidos sobre los corpora CoNLL 2002 y AnCora en el idioma español.

5.1. Metodología

5.1.1. Conjunto de datos

5.1.1.1. Corpora para NER

Para realizar evaluación en NER, mayormente se ha utilizado el corpus estándar CoNLL-2002 en español para establecer el estado del arte y realizar comparaciones entre los estudios realizados (Tjong Kim Sang, 2002; Tjong Kim Sang y De Meulder, 2003; dos Santos y Guimarães, 2015; Murthy y Bhattacharyya, 2016; Gillick et al., 2015; Lample et al., 2016; Yang et al., 2016). Sin embargo, CoNLL-2002 no es el único corpus anotado con entidades nombradas, el corpus AnCora (Recasens, 2008; Recasens et al., 2010) contiene entidades nombradas anidadas en el español.

Corpus CoNLL 2002 La tarea compartida CoNLL 2002 (Tjong Kim Sang, 2002) dio lugar a un corpus de entrenamiento y evaluación estándar para los algoritmos de NER supervisado: el corpus CoNLL-2002 español y holandés. CoNLL para el español está etiquetado con cuatro entidades: *Person*, *Organization*, *Location*, *Miscellaneous* y nueve clases (dado que se utiliza el estilo de anotación BIO, ver Sección 2.2 para explicación con mayor detalle): B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, B-MISC, I-MISC y O. En este corpus hay 74 683 *tokens* y 11 755 oraciones, un mayor detalle se observa en el Cuadro 5.1, asimismo, en el Cuadro 5.2 se aprecia el detalle de

las entidades anotadas en cada parte del corpus.

	Entrenamiento	Desarrollo	Prueba
Tokens	18 798	4 352	51 533
Oraciones	8 323	1 915	1 517

Cuadro 5.1: Cantidad de *tokens* y oraciones en CoNLL-2002 para el español

	Entrenamiento	Desarrollo	Prueba
LOC	4 914	985	1 084
MISC	2 173	445	340
ORG	7 390	1 700	1 400
PER	4 321	1 222	735
Total	18 798	4 352	3 559

Cuadro 5.2: Cantidad de entidades en CoNLL-2002 para el español

Corpus AnCora El corpus AnCora (para los idiomas catalán y español) está compuesto por anotaciones en diferentes niveles lingüísticos (*part-of-speech*, lemas, constituyentes, roles, clases semánticas de verbos, entidades nombradas, entre otras) (Recasens, 2008). El proceso de anotación ha sido llevado a cabo manualmente, semi automáticamente y automáticamente dependiendo del nivel lingüístico. En particular, las entidades nombradas fueron anotadas manualmente y contienen entidades anidadas. Este corpus tiene seis entidades: *Date*, *Location*, *Number*, *Organization*, *Person* y *Other*. Así también como el corpus CoNLL, en este corpus se hace uso del estilo de anotación BIO para entidades, por tanto, hay trece clases: B-DATE, I-DATE, B-LOCATION, I-LOCATION, B-NUMBER, I-NUMBER, B-ORGANIZATION, I-ORGANIZATION, B-OTHER, I-OTHER, B-PERSON, I-PERSON y O.

Asimismo, se ha utilizado la versión del corpus procesada por la tarea compartida SemEval-2010 (Recasens et al., 2010) enfocada en *Coreference Resolution*. Se ha utilizado esta versión del corpus debido a que proporciona un conjunto de entrenamiento, desarrollo y prueba a diferencia del original AnCora que no lo proporciona. Un último procesamiento llevado a cabo fue eliminar las entidades anidadas, manteniendo las entidades en primer nivel (entidad que agrupaba a las demás) dado que el tratamiento de entidades anidadas no es parte de este estudio.

Luego, en este corpus hay 20 633 *tokens* y 12 146 oraciones (mayor detalle en el Cuadro 5.3), asimismo, en el Cuadro 5.4 se aprecia el detalle de las entidades anotadas en cada parte del corpus.

	Entrenamiento	Desarrollo	Prueba
Tokens	15 546	2 325	2 762
Oraciones	9 022	1 419	1 705

Cuadro 5.3: Cantidad de *tokens* y oraciones en AnCora para el español

	Entrenamiento	Desarrollo	Prueba
DATE	1 185	985	1 084
LOC	2 339	332	416
NUMBER	1 227	187	270
ORG	4 528	726	821
OTHER	1 353	184	216
PERSON	4 914	685	823
Total	15 546	3 099	3 630

Cuadro 5.4: Cantidad de entidades en AnCora para el español

5.1.1.2. Texto sin anotación

El texto sin anotación ha sido utilizado para aprender características de forma no supervisada de las palabras contenidas en el texto, por esta razón se utiliza un conjunto de datos en español y un segundo conjunto en inglés.

Conjunto de datos en Español Con el fin de calcular las *word representations* (*Brown clusters*, *word embeddings*) y colocaciones para el idioma español se requiere de una gran cantidad de datos no etiquetados. Para este fin se ha utilizado el corpus y *embeddings* de *Spanish Billion Words* (SBW) (Cardellino, 2016). Este conjunto de datos fue reunido de varias fuentes de dominio público [1] en español. Entre ellas: Tibidabo Treebank, IULA Spanish LSP Treebank, conjunto de datos de Wikipedia en español, Wikisource, Wikibooks hasta setiembre del 2015 y la porción en español de SenSem, AnCora, Europarl y el proyecto OPUS (Cardellino, 2016). Este conjunto de datos contiene 3 817 833 *tokens únicos*, y 1 000 653 *tokens únicos* de embeddings con 300 dimensiones por vector.

Conjunto de datos en Inglés Los nombres de las entidades tienden a ser muy similares en diversos idiomas y dominios. Esto implica que *word representations* deberían incrementar el rendimiento en el reconocimiento de las entidades cuando se utilizan conjuntos de datos multilenguaje. Para esto, se ha utilizado un corpus de Wikipedia en Inglés hasta el 2012 preprocesado por Guo et al. (Guo et al., 2014). En este conjunto de datos los párrafos que contenían carácteres no romanos fueron eliminados, las palabras fueron convertidas a minúsculas y se eliminaron las palabras menos frecuentes.

[1] http://crscardellino.me/SBWCE/

5.1.2. Detalles de implementación

A continuación se describen algunas pautas en el proceso de implementación.

- Se utilizó CRF de cadena lineal y las características de línea base empleadas han sido definidas en la Sección 4.2.
- El estilo de anotación empleado es BIO (definido en la Sección 2.2).
- Las características (supervisadas y no supervisadas) son generadas para cada *token* en una etapa previa al entrenamiento, para lo cual es necesario haber procesado los conjuntos de datos sin anotación, para tener los *word representations* y colocaciones.
- Con respecto a los *embeddings* binarizados, luego de obtener el 'vector binarizado', cada componente del vector es utilizada como característica siempre y cuando el valor binarizado sea diferente de cero.
- Con respecto a *Brown clustering*, se definió el número k de clusters a 1000, de acuerdo a Turian et al. (Turian et al., 2010).
- Para modelar diferentes niveles de granularidad en los clustering de *embeddings* se han calculado 500, 1000, 1500, 2000 y 3000 clusters (Guo et al., 2014).
- El número de prototipos distribucionales empleados en cada corpus fue definido en cada conjunto de desarrollo (*development set*). Al evaluar sobre el corpus CoNLL-2002 empleando los primeros 40 prototipos se alcanzó un *F1* de 75,79 % (como se aprecia en el Cuadro 5.5a), mientras que para AnCora se empleó 80 prototipos alcanzando un *F1* de 61,17 % (ver Cuadro 5.5b). Finalmente, estos son los números de prototipos empleados en cada corpus con los que se alcanzaron los resultados reportados en este trabajo.

Nro. Prototipos	***F1***
20	75.64 %
40	**75.79 %**
80	75.66 %
100	75.52 %

(a) Número de prototipos en CoNLL-2002

Nro. Prototipos	***F1***
40	61.01 %
60	61.12 %
80	**61.17 %**
100	61.15 %

(b) Número de prototipos en AnCora

Cuadro 5.5: Número de prototipos por cada corpus

- Durante el proceso del calculo de los prototipos, en la etapa del calculo de similitud de cada palabra (de SBW) con el conjunto de *prototipos*, fue necesario utilizar paralelismo en CPU, dado que aproximadamente eran 1 billón de palabras con las que se tenia que buscar los prototipos (los k más representativos por cada clase) más similares.

- El lenguaje de programación empleado fue Python, dado que nos permitió manejar de forma directa la generación de características y el tratamiento del texto de entrada.

5.2. Resultados

En esta sección se presentan los resultados. Primero, se presentan los experimentos preliminares realizados en HMM y Polyglot-NER, con fines de comparación. Seguidamente, los experimentos y resultados alcanzados con la propuesta (evaluados en los corpora CoNLL-2002 y AnCora en el idioma español). Finalmente, una discusión de los resultados alcanzados (comparando la métrica *F1* alcanzada y tiempos de entrenamiento por los modelos de aprendizaje en el estado del arte).

5.2.1. Experimentos preliminares: HMM y Polyglot-NER

En esta subsección se presentan los experimentos realizados, para efectos de comparación con los enfoques de clasificación de secuencias. Primero, se experimenta con HMM y luego con el modelo entrenado de Polyglot-NER (Al-Rfou et al., 2015).

5.2.1.1. HMM

Un enfoque probabilístico comúnmente empleado para tratar la clasificación de secuencias es HMM (Florian et al., 2003). Se ha construido un etiquetador de entidades basado en las distribución de probabilidad definida en la Sección 2.3.1 y se ha llevado a cabo un conjunto de experimentos aplicados directamente al modelo, la tendencia de los resultados alcanzados se muestran en la Figura 5.1 observando que a partir del experimento número ocho los resultados tienden a ser estables.

En la Figura 5.1 se muestran los resultados alcanzados (eje Y) por cada experimento (eje X). En cada experimento se consideró diferentes valores para los parámetros λ_i de la Ecuación 2.8, evaluado en el conjunto de prueba de CoNLL-2002. El detalle de los mejores resultados aplicando HMM en CoNLL-2002 se encuentran en el Cuadro 5.6 con $\lambda_1 = 0{,}1$, $\lambda_2 = 0{,}9$ y $\lambda_3 = 0{,}0$ de la Ecuación 2.8. Se alcanzó un $F1$ de 60,13 %.

Con la distribución de probabilidad definida en la Sección 2.3.1, Ecuación 2.13, tomando en cuenta la característica del uso de mayúscula, se realizaron otro grupo de

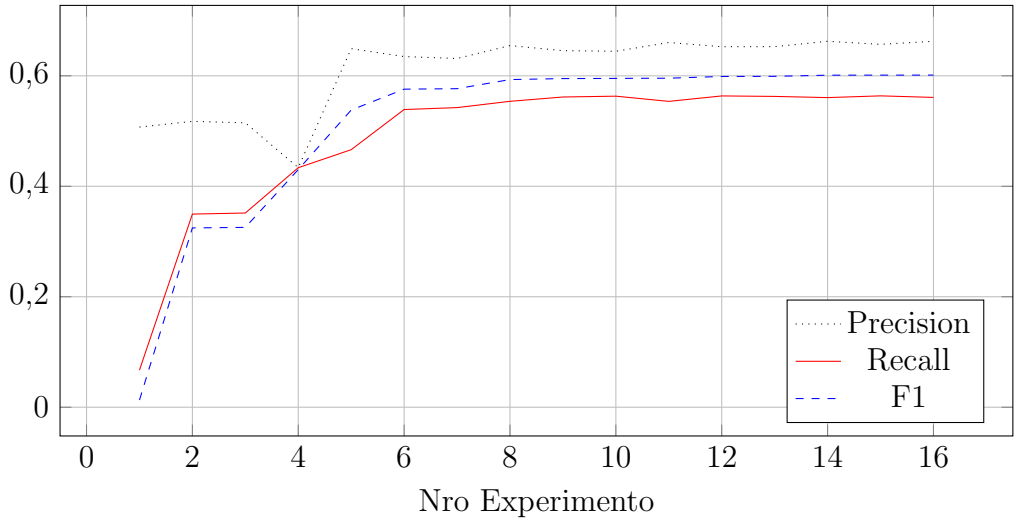


Figura 5.1: Experimentos con HMM en CoNLL-2002

	Precision	**Recall**	**F1**
Location	56.01 %	61.43 %	58.60 %
Miscellaneous	53.81 %	35.29 %	42.62 %
Organization	68.55 %	65.85 %	67.17 %
Person	86.64 %	61.76 %	72.12 %
Total	62.25 %	56.08 %	60.13 %

Cuadro 5.6: HMM evaluado sobre CoNLL-2002 español

experimentos con una tendencia mostrada en la Figura 5.2, notando estabilidad en la curva *F1* a partir del experimento número 10, así también como en *Precision* y *Recall*.

De forma análoga al conjunto de experimentos anterior, en la Figura 5.2 se muestran los resultados alcanzados (eje Y) por cada experimento (eje X). En cada experimento se consideró diferentes valores para los parámetros λ_i y para γ_i de la Ecuación 2.13, evaluado en el conjunto de prueba de CoNLL-2002. Los resultados con mayor *F1* alcanzan 60.24 % y se detallan en el Cuadro 5.7 donde observamos pequeñas mejoras en el rendimiento del clasificador. Asimismo, en este experimento $\lambda_1 = 0{,}3$, $\lambda_2 = 0{,}6$, $\lambda_3 = 0$, $\lambda_4 = 0{,}05$, $\lambda_5 = 0{,}05$, $\gamma_1 = 0{,}8$ y $\gamma_2 = 0{,}2$.

	Precision	**Recall**	**F1**
Location	56.74 %	63.65 %	60.00 %
Miscellaneous	52.67 %	34.70 %	41.84 %
Organization	69.53 %	65.85 %	67.64 %
Person	84.85 %	61.76 %	71.49 %
Total	65.95 %	56.49 %	60.24 %

Cuadro 5.7: HMM con una característica evaluado sobre CoNLL-2002 español

Tanto para el caso de la aplicación de HMM directamente (ver Cuadro 5.6) como

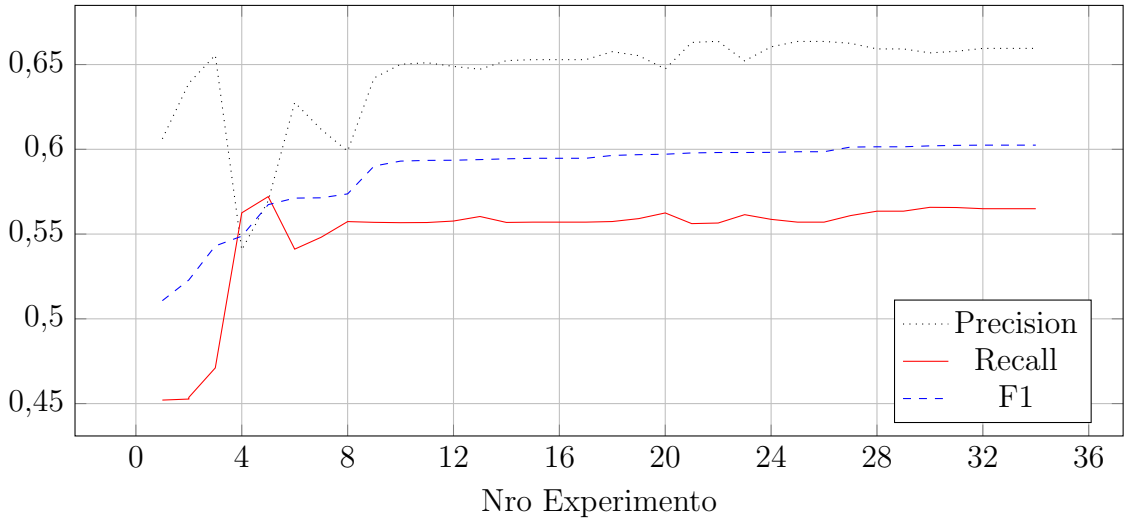


Figura 5.2: Experimentos con HMM utilizando *mayúsculas* como característica en CoNLL-2002

en HMM con una característica (ver Cuadro 5.7), los resultados con más alto resultado no tienen en cuenta los unigramas por el contrario si a los bigramas.

5.2.1.2. Polyglot-NER

El estudio llevado a cabo en Polyglot-NER [2] (Al-Rfou et al., 2015) dió como resultado un modelo entrenado con las entidades aprendidas haciendo uso de Wikipedia y Freebase, el cual ha sido utilizado en este trabajo para evaluar la tarea sobre el corpus CoNLL-2002 en español. Polyglot-NER utiliza el estilo de anotación IO y solamente tres entidades nombradas, por lo que al experimentar con el conjunto de evaluación de CoNLL-2002 (el cual tiene cuatro entidades) hubo una entidad que mostraba resultados cero (ver Cuadro 5.8).

	Precision	**Recall**	**F1**
Location	58.50 %	74.26 %	65.45 %
Miscellaneous	0.00 %	0.00 %	0.00 %
Organization	61.01 %	38.79 %	47.42 %
Person	71.40 %	87.62 %	78.68 %
Total	62.88 %	55.97 %	59.22 %

Cuadro 5.8: Polyglot-NER evaluado sobre CoNLL-2002 español en estilo IO

A continuación se describirá el procesamiento realizado a los resultados de Polyglot-NER es con fin de conseguir resultados adecuados con respecto a la salida de la herramienta y el corpus de evaluación que es CoNLL-2002.

[2] http://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html

El primer procesamiento posterior a los resultados obtenidos por Polyglot-NER fue convertir el estilo de anotación de IO a BIO (CoNLL está en este estilo de anotación) obteniendo los resultados del Cuadro 5.9. La conversión del estilo IO a BIO no impactó de forma global al resultado de 59.22 % pero si lo hizo en favor del *recall* y contra la *precision*, mínimamente. Luego se aprecia que la entidad *Person* es beneficiada con esta conversión.

	Precision	**Recall**	**F1**
Location	58.49 %	74.35 %	65.48 %
Miscellaneous	0.00 %	0.00 %	0.00 %
Organization	61.01 %	38.79 %	47.42 %
Person	71.24 %	87.62 %	78.58 %
Total	62.83 %	56.00 %	59.22 %

Cuadro 5.9: Polyglot-NER evaluado sobre CoNLL-2002 español en estilo BIO

Como paso final en el procesamiento, no se toma en cuenta la entidad *Miscellaneous* en el corpus CoNLL-2002, ya que no está presente en Polyglot, así que esa entidad es eliminada del corpus (se cambian las etiquetas de esta entidad por la etiqueta de *no entidad*). El resultado se aprecia en el Cuadro 5.10 donde hay un incremento mayor en el rendimiento global del clasificador de Polyglot-NER entrenado con Wikipedia en español (entidades aprendidas de forma no supervisada) y evaluado con CoNLL-2002 alcanzando un 62.37 % de *F1*.

	Precision	**Recall**	**F1**
Location	58.49 %	74.35 %	65.48 %
Organization	61.01 %	38.79 %	47.42 %
Person	71.24 %	87.62 %	78.58 %
Total	62.83 %	61.91 %	62.37 %

Cuadro 5.10: Polyglot-NER evaluado sobre CoNLL-2002 español en estilo BIO sin entidad *Miscellaneous*

5.2.2. Experimentos y resultados de la propuesta

En el Cuadro 5.11 se muestra un resumen de los experimentos previamente mostrados con el fin de realizar comparaciones con el modelo propuesto, junto con el estado del arte para NER en el idioma español (dos Santos y Guimarães, 2015; Gillick et al., 2015; Murthy y Bhattacharyya, 2016; Lample et al., 2016; Yang et al., 2016). Para evaluar los modelos se ha utilizado el script estándar `conlleval`[3].

Dado que CRF permite modelar conocimiento en el modelo (aprendido de forma supervisada o no supervisada) a diferencia de HMM, esto se refleja en los resultados donde en el Cuadro 5.11, HMM alcanza un *F1* de 60.24 % y la línea base de este trabajo

[3] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt

un 80.02 %. También se aprecia que Polyglot alcanza un 62.37 % de *F1* pero con un conjunto de entrenamiento mucho mayor, ya que proviene de entidades aprendidas de Wikipedia, pero aún así no alcanza a superar nuestra línea base.

Modelo	**F1**
HMM	60.24 %
Polyglot (Al-Rfou et al., 2015)	62.37 %
Carreras (Carreras et al., 2002)*	79.28 %
Línea base de este trabajo	**80.02 %**
Carreras (Carreras et al., 2002)	81.39 %
Stanford CRF NER (Finkel et al., 2005)**	81.02 %
Stanford CRF NER (Finkel et al., 2005)	81.44 %
CNN para palabras y caracteres (dos Santos y Guimarães, 2015)	82.21 %
CNN+RNN (Murthy y Bhattacharyya, 2016)	82.59 %
RNN (Gillick et al., 2015)	82.95 %
RNN+CRF (Lample et al., 2016)	85.75 %
Gated RNN+CRF (Yang et al., 2016)	85.77 %

Cuadro 5.11: Resumen de experimentos en CoNLL-2002 español

*Sin tomar en cuenta gazetteers.
**Haciendo uso de características supervisadas.

Los resultados alcanzados por el modelo propuesto muestran una tendencia en la Figura 5.3. Se puede apreciar los resultados alcanzados (eje Y) por cada experimento (eje X). En cada experimento se han considerado diferentes características, encontrando que la combinación del uso de las características no supervisadas dadas por *Brown clustering* (calculados del conjunto de datos en inglés), *clustering embeddings* y prototipos alcanzaron los mejores resultados con un *F1* de 82.44 %, siendo evaluados sobre el conjunto de prueba de CoNLL-2002. El detalle de los resultados alcanzados en los experimentos se muestra en el Cuadro 5.12.

Cabe señalar que *Brown clustering* mejora la línea base así también como el enfoque de las *colocaciones* y *clustering embeddings*. En contraste, el enfoque de *embeddings binarizados* presenta un rendimento menor que la línea base. Esto parece ser debido al hecho que en los *embeddings* binarizados al agrupar las dimensiones en un conjunto finito de valores discretos sesga información relevante para NER en español.

Lo mismo ocurre en *Prototipos*, los cuales cuando se aplican directamente resultan ser de menor rendimiento que la línea base.

Al combinar los enfoques de características no supervisadas, se llega a alcanzar y superar la línea base, así como *Brown clustering* y *clustering embeddings* aplicados directamente.

Sin embargo, los mejores resultados se obtuvieron utilizando una combinación *multi-lenguaje* de *Brown clusters* calculados del conjunto de datos de Wikipedia en inglés (2012) con *clustering embeddings* y *prototipos* calculados de SBW. Lo mismo se

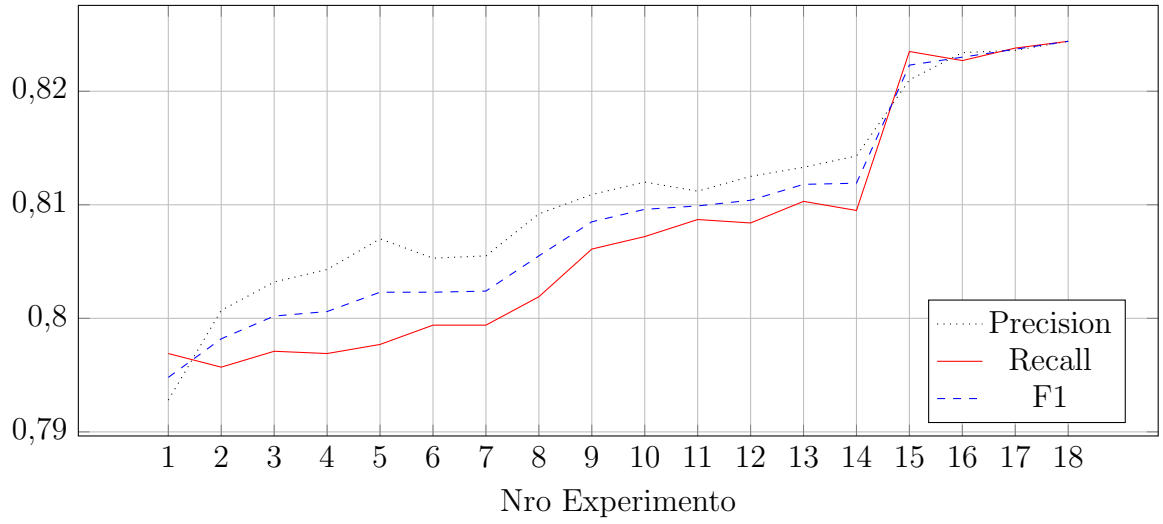


Figura 5.3: Enfoques propuestos en CRF para NER en CoNLL-2002

mantiene al combinar *Brown clusters*, *clustering* de *embeddings*, prototipos y colocaciones.

La razón por la que los *Brown clusters* han impactado positivamente en esta tarea se debe al alto grado de solapamiento a lo largo de las entidades entre el español y el inglés. Dicho de otra forma, muchas entidades comparten los mismos nombres y ocurren en contextos similares para ambos idiomas, lo que conlleva a obtener características con alto nivel predictivo.

El segundo conjunto de datos utilizado para evaluar NER en el idioma español es AnCora, la tendencia de los resultados obtenidos se muestra en la Figura 5.4, donde los resultados alcanzados se muestran en el eje Y por cada experimento (eje X). De forma análoga al grupo de experimentos anteriores, en cada experimento se han considerado diferentes características, encontrando que la mejor combinación de características no supervisadas alcanzaron los mejores resultados con un *F1* de 65.72 %, siendo evaluados sobre el conjunto de prueba del corpus. El detalle de los resultados alcanzados en los experimentos se muestra en el Cuadro 5.13. Con el fin de realizar una comparación en este *corpus* se ha evaluado el mismo en Stanford CRF NER.

La evaluación en AnCora muestra que todos los enfoques superan la línea base (se emplearon las características definidas en la Sección 4.2), a excepción de la combinación de *prototipos* y *colocaciones*. Cabe señalar que el enfoque *clustering embeddings* muestra alto rendimiento con respecto a la línea base y los resultados alcanzados empleando Stanford CRF NER (Finkel et al., 2005). En contraste con el corpus CoNLL, en AnCora, el uso de *colocaciones* en combinación con *Brown clustering* (calculados del idioma inglés), *clustering embeddings* y *prototipos* han dado lugar a los mejores resultados.

Modelo	Precision	Recall	F1
Línea base	80.32 %	79.71 %	80.02 %
+Binarización	79.28 %	79.69 %	79.48 %
+*Brown*	80.12 %	80.87 %	80.99 %
+Prototipo	80.07 %	79.57 %	79.82 %
+Colocación	80.70 %	79.77 %	80.23 %
+*Clustering*	80.55 %	79.94 %	80.24 %
+*Clustering*+Prototipo	80.92 %	80.19 %	80.55 %
+*Brown*+Colocación	81.25 %	80.84 %	81.04 %
+*Brown*+*Clustering*	81.33 %	81.03 %	81.18 %
+*Brown*+*Clustering*[*]	82.34 %	82.27 %	82.30 %
+*Brown*+*Clustering*+Prototipo	81.43 %	80.95 %	81.19 %
+*Brown*+*Clustering*+Prototipo+Colocación	81.20 %	80.72 %	80.96 %
+*Brown*+*Clustering*+Prototipo+Colocación*	82.10 %	82.35 %	82.23 %
+*Brown*+*Clustering*+Prototipo*	**82.44 %**	**82.44 %**	**82.44 %**

Cuadro 5.12: Resultados en CoNLL-2002 para el español con características no supervisadas en CRF

.

[*] *Brown clusters* en idioma inglés

Discusión Los primeros resultados para NER en español utilizando el corpus CoNLL 2002, sin considerar conocimiento externo alcanza un 79.28 % de *F1* mientras que haciendo uso de *gazetteers* se consiguió un 81.39 % de *F1*.

Cabe señalar, que el conocimiento introducido a modelos de aprendizaje previos fue *supervisado*. Por otro lado, en este trabajo se ha considerado conocimiento externo *no supervisado* que mejora significativamente los resultados. Esto es corroborado por la exploración de características no supervisadas con Stanford NER CRF (Finkel et al., 2005) alcanzando 81.44 %, nuevamente por encima de Carreras et al. (Carreras et al., 2002) lo cual incide en el uso de características no supervisadas para mejorar el rendimiento del etiquetador.

Lo que es más importante, esta propuesta ha demostrado que un recurso en inglés (*Brown clusters* calculados de Wikipedia en inglés) puede ser utilizado para mejorar NER en español con *word representations* dado que:

1. Entidades en inglés y español tienen muchas similitudes.

2. *Brown clusters* calculados para entidades en inglés, correlaciona mejor dando lugar a un mejor modelo.

Otro punto a considerar, es que mientras el enfoque de binarización mejora la línea base para NER en inglés (Guo et al., 2014), no necesariamente pasa lo mismo con el español. Al parecer este enfoque añade ruido a NER en español. Así como la combinación con *colocaciones* no alcanza tan buenos resultados en CoNLL-2002 ya

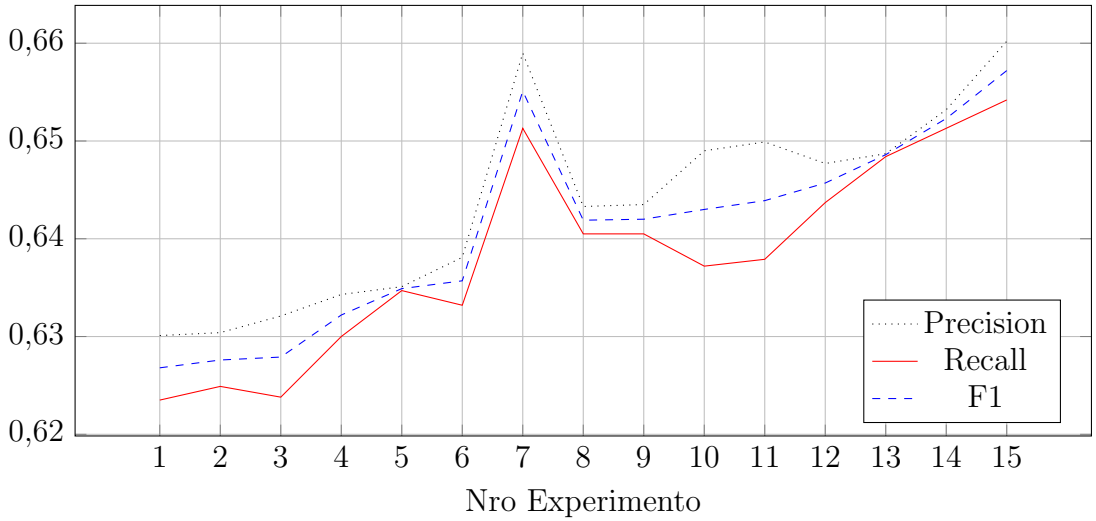


Figura 5.4: Enfoques propuestos en CRF para NER en AnCora

que las entidades no están conformadas por colocaciones mientras que en AnCora este fenómeno si está presente.

También se ha notado que el uso original de *mayúsculas/minúsculas* tiene un impacto diferente en el trabajo propuesto. Con la siguiente configuración: *Brown clusters* en inglés, *clustering embeddings* en español y *prototipos* en español convertidos a *minúsculas*, se consigue 81.6 % de *F1*, mientras que bajo la misma configuración pero con los prototipos en español con el uso de mayúsculas original se consigue 82.38 %, es decir que el uso de las mayúsculas causó un 0.78 % de variación, utilizando *word embeddings* del conjunto de desarrollo de CoNLL. Mientras que al realizar esta comparación con un conjunto de *word embeddings* ampliamente mayor proveniente de SBW, la diferencia se hace menor. Este último hecho indica que entre mayor sea el vocabulario de los *embeddings* utilizados, menor es el impacto del uso original de *mayúsculas/minúsculas*.

Analizando el hecho del uso original de *mayúsculas/minúsculas* se ha encontrado que se debe a que con los prototipos en minúsculas se ignora el contexto real en el cual la palabra aparece (dado que un prototipo es un ejemplo de una clase) y por tanto serán mapeados a un vector de palabra incorrecto en los *embedding* (al momento de calcular la similaridad del coseno).

Emplear colocaciones como características puede sugerir que proporcionan información complementaria para NER, como se aprecia en el Cuadro 5.12. Sin embargo, al realizar combinaciones con otros enfoques, las colocaciones añaden ruido para el caso de CoNLL-2002.

Luego, comparando el trabajo propuesto con el estado del arte actual (véase Cuadro 5.11), basado en la utilización de métodos de *Deep Learning* (dos Santos y Guimarães, 2015; Gillick et al., 2015; Murthy y Bhattacharyya, 2016; Lample et al., 2016; Yang et al., 2016) (que extraen características a nivel de carácter, palabra y bytes para aprender sus modelos), se supera en términos de *F1* el trabajo de dos Santos y

Modelo	Precision	Recall	FB1
Línea base	63.04 %	62.35 %	62.76 %
+Brown	63.51 %	63.47 %	63.49 %
+Prototipo	63.43 %	63.00 %	63.22 %
+Colocación	63.21 %	62.38 %	62.79 %
+*Clustering*	65.32 %	65.13 %	65.23 %
+*Clustering*+Prototipo	64.87 %	64.84 %	64.86 %
+*Brown*+*Clustering*	64.77 %	64.37 %	64.57 %
+*Clustering*+Colocación	64.35 %	64.05 %	64.20 %
+*Brown*+Colocación	63.81 %	63.32 %	63.57 %
+Prototipo+Colocación	63.01 %	62.35 %	62.68 %
+*Brown*+*Clustering*+Prototipo*	65.90 %	65.13 %	65.51 %
+*Brown*+*Clustering*+Prototipo	64.33 %	64.05 %	64.19 %
+*Brown*+*Clustering*+Colocación	64.90 %	63.72 %	64.30 %
+*Brown*+*Clustering*+Prototipo+Colocación	64.99 %	63.79 %	64.39 %
+*Brown*+*Clustering*+Prototipo+Colocación*	**66.02 %**	**65.42 %**	**65.72 %**
Finkel (Finkel et al., 2005)	63.67 %	60.10 %	61.84 %
Finkel** (Finkel et al., 2005)	63.04 %	62.35 %	62.36 %

Cuadro 5.13: Resultados en AnCora para el español.

Brown clusters en idioma inglés.
**Usando una característica no supervisada.

Guimarães (dos Santos y Guimarães, 2015) y se acerca a Murthy y Bhattacharyya (Murthy y Bhattacharyya, 2016) y a Gillick et al. (Gillick et al., 2015).

Los tiempos de ejecución de los experimentos realizados y las propuestas del estado del arte se muestran en el Cuadro 5.14[4]. Los resultados con mayor *F1* de la propuesta de este trabajo tomó 16 minutos en realizar el entrenamiento, lo cual es un tiempo mucho menor en relación a los otros estudios, incluso de dos Santos y Guimarães (dos Santos y Guimarães, 2015), considerando que la propuesta de este trabajo lo supera en *F1*. A consecuencia de este hecho, la ventaja es que podemos aprovechar el uso de recursos no supervisado de mayor tamaño al actual.

Experimentos adicionales en el corpus AnCora confirman que el uso de *word representations* multilenguaje brinda información complementaria para reconocer entidades (incluso cuando hay entidades anidadas). Como se puede apreciar en el Cuadro 5.13 la mejor combinación alcanzó 65.72 % de *F1*. Esto último se debe a que las entidades anidadas en este corpus en muchos casos están compuestas por colocaciones.

Finalmente, en el Cuadro 5.15a y el Cuadro 5.15b se aprecia el resumen de los experimentos para NER en el idioma español. En la parte superior del Cuadro 5.15a se presentan los resultados alcanzados por este trabajo en el corpus CoNLL-2002, mientras que en la parte inferior se muestra el *F1* de otros modelos de aprendizaje, incluyendo

[4]Los tiempos de ejecución de los trabajos del estado del arte fueron consultados a los autores de las respectivas propuestas

Modelo	**F1**	**Tiempo de ejecución**	**UP** *
Línea base+*Brown*+*Clustering*+Prototipo**	82.44 %	16 min	CPU
Stanford CRF NER (Finkel et al., 2005)	81.44 %	46 min	CPU
CNN (dos Santos y Guimarães, 2015)	82.21 %	38 min	CPU(2 hilos)
CNN+RNN (Murthy y Bhattacharyya, 2016)	82.59 %	8h	GPU
RNN (Gillick et al., 2015)	82.95 %	días	CPU
RNN+CRF (Lample et al., 2016)	85.75 %	9h	GPU
Gated RNN+CRF (Yang et al., 2016)	85.77 %	10h	GPU

Cuadro 5.14: Tiempos de ejecución en entrenamiento sobre CoNLL-2002

*Unidad de Procesamiento.
**_Brown clusters_ en idioma inglés.

del estado del arte. De forma similar se muestra en el Cuadro 5.15b, pero solamente fue evaluado con Stanford CRF NER.

5.3. Consideraciones Finales

En este capítulo se ha presentado la metodología de evaluación mediante el conjunto de datos empleado, detalles de implementación y finalmente los resultados que se alcanzan con la propuesta.

Modelo	**F1**
Línea base	80.02 %
+*Clustering*+Prototipo	80.55 %
+*Brown*+Colocación	81.04 %
+*Brown*+*Clustering*	81.18 %
+*Brown*+*Clustering*[*]	82.30 %
+*Brown*+*Clustering*+Prototipo	81.19 %
+*Brown*+*Clustering*+Prototipo+Colocación	80.96 %
+*Brown*+*Clustering*+Prototipo+Colocación*	82.23 %
+Brown+Clustering+Prototipo*	**82.44 %**
HMM	60.24 %
Polyglot (Al-Rfou et al., 2015)	62.37 %
Carreras (Carreras et al., 2002)[**]	79.28 %
Carreras (Carreras et al., 2002)	81.39 %
Stanford CRF NER (Finkel et al., 2005)[***]	81.02 %
Stanford CRF NER (Finkel et al., 2005)	81.44 %
CNN (dos Santos y Guimarães, 2015)	82.21 %
CNN+RNN (Murthy y Bhattacharyya, 2016)	82.59 %
RNN (Gillick et al., 2015)	82.95 %
RNN+CRF (Lample et al., 2016)	85.75 %
Gated RNN+CRF (Yang et al., 2016)	85.77 %

(a) Resultados en CoNLL-2002

[*] *Brown clusters* en idioma inglés.
[**] Sin tomar en cuenta gazetteers.
[***] Haciendo uso de características supervisadas.

Modelo	**F1**
Línea base	62.76 %
+*Clustering*+Prototipo	64.86 %
+*Brown*+*Clustering*	64.57 %
+*Clustering*+Colocación	64.20 %
+*Brown*+Colocación	63.57 %
+Prototipo+Colocación	62.68 %
+*Brown*+*Clustering*+Prototipo[*]	64.19 %
+*Brown*+*Clustering*+Prototipo	64.19 %
+*Brown*+*Clustering*+Colocación	64.30 %
+*Brown*+*Clustering*+Prototipo+Colocación	64.39 %
+Brown+Clustering+Prototipo+Colocación*	**65.72 %**
Finkel (Finkel et al., 2005)	61.84 %
Finkel[**] (Finkel et al., 2005)	62.36 %

(b) Resultados en AnCora

[*] *Brown clusters* en idioma inglés.
[**] Usando una característica no supervisada.

Cuadro 5.15: Resumen de resultados en NER para el idioma español

Capítulo 6

Conclusiones y Trabajos Futuros

Las principales conclusiones a las que se ha llegado en este trabajo son:

- En este trabajo se ha estudiado el reconocimiento de entidades nombradas para el idioma español en un modelo de aprendizaje de secuencias de texto, con énfasis en el uso de características no supervisadas. Se identificó las condiciones en las que se mejora el *F1* en la tarea.

- Se ha explorado características supervisadas (de línea base) y no supervisadas. Estas últimas basadas por lo general en *word representations* multilenguaje, dentro de un modelo de clasificación CRF para NER en español, entrenado sobre el corpus CoNLL-2002 y AnCora en español, *Spanish Billion Word* y Wikipedia en inglés (conjunto de datos hasta el 2012). Esta es una nueva propuesta para el idioma español. Los experimentos muestran resultados competitivos al compararlos con el estado del arte actual de NER en español, basado en *Deep Learning*. En particular, en este trabajo se supera a dos Santos y Guimarães (dos Santos y Guimarães, 2015).

- Con respecto a las técnicas de aprendizaje profundo que actualmente establecen el estado del arte en la tarea, se ha presentado una comparación de los tiempos de ejecución de las propuestas, siendo más conveniente aplicar la propuesta cuando no se dispone del tiempo suficiente o de recursos computacionales necesarios para entrenar estos modelos, a un costo de un menor *F1*, sin embargo competitivo. Por otro lado, aplicando esta propuesta se tiene la oportunidad de incrementar el tamaño del conjunto de datos no supervisado con el fin de mejorar la calidad de las características no supervisadas aprendidas.

- *Word representations* multilenguaje como características en un modelo CRF tienen un impacto positivo en el rendimiento de NER para el idioma español probado sobre dos diferentes corpora. Mientras que las *colocaciones* han sido mayormente de impacto positivo cuando las entidades a ser nombradas las contienen, de lo contrario esta característica provoca un menor rendimiento.

- Se recomienda emplear un conjunto de texto de gran tamaño (como SBW) para aprender las características no supervisadas, en particular para el cálculo de los *word embeddings* dado que ello contribuye a ponderar cada *token* en diferentes contextos. De la misma forma, se ha mostrado que al usar *word embeddings* calculados de texto de mayor tamaño, se atenúa el impacto del uso de *mayúsculas/minúsculas* en el proceso de cálculo de *prototipos*.

6.1. Limitaciones

Este trabajo contempla el estudio del reconocimiento de entidades nombradas en el idioma español, por ello se ha evaluado sobre el corpus estandarizado CoNLL-2002 y AnCora (ambos para el idioma estudiado), sin entrar en mayor detalle de granularidad de entidades y de niveles de anidación. Por un lado, se estudian las entidades presentes en tales corpus (persona, organización, ubicación, fecha, número, otros) y no se consideran entidades como profesiones, que es un grado de granularidad de entidades mayor. Por otro lado, con respecto al nivel de anidado no se ha estudiado entidades anidadas (dicho de otra manera, entidades contenidas en otras).

6.2. Recomendaciones

Se han realizado estudios que muestran que utilizando más dimensiones para los *word embeddings* resulta en distinguir características obtenidas de las palabras, por ello se recomienda explorar el impacto del número de dimensiones utilizadas para el cálculo de los *word embeddings* en esta tarea para el idioma español.

6.3. Trabajos futuros

En el futuro, se espera explotar aún más el uso de recursos *multilenguaje* y examinar conjuntos de datos (a gran escala). Asimismo, explorar la importancia de la familia del idioma para elegir un recurso *multilenguaje* como recurso adicional.

En este trabajo se presentan experimentos y resultados relevantes para el estudio de NER en el idioma español, otro trabajo futuro surge en la aplicación de las características no supervisadas bajo los enfoques explicados en otras tareas que aprovechen el conjunto de datos sin anotaciones que es ampliamente disponible.

Con el fin de hacer esta propuesta escalable se sugieren dos opciones. Primero, calcular el proceso de prototipos en GPU, ya que es el que más tiempo toma (actualmente se usa hilos en CPU). Como segunda opción se sugiere emplear CRF con hilos en CPU y posteriormente en GPU.

Bibliografía

Aggarwal, C. y Zhai, C. (2012). *Mining Text Data*. Springer.

Al-Rfou, R., Kulkarni, V., et al. (2015). Polyglot-NER: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30 - May 2, 2015*.

Barceló, G., Cendejas, E., et al. (2009). *Computational Linguistics and Intelligent Text Processing: 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1-7, 2009. Proceedings*, chapter Formal Grammar for Hispanic Named Entities Analysis, pages 183–194. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bingel, J. y Haider, T. (2014). Named Entity Tagging a Very Large Unbalanced Corpus: Training and Evaluating NE Classifiers. In Calzolari, N., Choukri, K., et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In Chiarcos, C., de Castilho, E., et al., editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, pages 31–40, Tübingen. Gunter Narr Verlag.

Brants, T. (2000). TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 224–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brown, P. F., deSouza, P. V., et al. (1992). Class-based N-gram Models of Natural Language. *Comput. Linguist.*, 18(4):467–479.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings.

Carreras, X., Màrques, L., et al. (2002). Named entity extraction using AdaBoost. In *Proceedings of CoNLL-2002*, pages 167–170. Taipei, Taiwan.

Carreras, X. y Màrquez, L. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Çelik, K. (2012). A comprehensive analysis of using Wordnet, part-of-speech tagging, and Word Sense Disambiguation in Text Categorization. Master's thesis, Institute for Graduate Studies in Science and Engineering.

Che, W., Wang, M., et al. (2013). Named entity recognition with bilingual constraints. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 52–62.

Collobert, R., Weston, J., et al. (2011). Natural Language Processing (almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Copara, J., Ochoa, J., et al. (2016a). *Conditional Random Fields for Spanish Named Entity Recognition Using Unsupervised Features*, pages 175–186. Springer International Publishing, Cham.

Copara, J., Ochoa, J., et al. (2016b). Spanish NER with Word Representations and Conditional Random Fields. In *Proceedings of the Sixth Named Entity Workshop*, pages 34–40, Berlin, Germany. Association for Computational Linguistics.

Copara, J., Ochoa, J., et al. (2016c). Exploring Unsupervised Features in Conditional Random Fields for Spanish Named Entity Recognition. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 283–288.

Deng, L. y Yu, D. (2014). Deep Learning: Methods and Applications. Technical Report MSR-TR-2014-21.

Dietterich, T. G. (2002). *Machine Learning for Sequential Data: A Review*, pages 15–30. Springer Berlin Heidelberg, Berlin, Heidelberg.

dos Santos, C. y Guimarães, V. (2015). Boosting Named Entity Recognition with Neural Character Embeddings. In *Proceedings of the Fifth Named Entity Workshop*, pages 25–33, Beijing, China. Association for Computational Linguistics.

Faruqui, M. y Padó, S. (2010). Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Finkel, J. R., Grenager, T., et al. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Finkel, J. R. y Manning, C. D. (2009). Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 141–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Florian, R., Ittycheriah, A., et al. (2003). Named Entity Recognition through Classifier Combination. In Daelemans, W. y Osborne, M., editors, *Proceedings of CoNLL-2003*, pages 168–171. Edmonton, Canada.

Galicia-Haro, S. N., Gelbukh, A., et al. (2004). *Recognition of Named Entities in Spanish Texts*, pages 420–429. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gillick, D., Brunk, C., et al. (2015). Multilingual Language Processing From Bytes. *ArXiv e-prints*.

Glavaš, G., Karan, M., et al. (2012). CroNER: A State-of-the-Art Named Entity Recognition and Classification for Croatian. In *Information Society 2012-Eighth Language Technologies Conference IS-LTC 2012*, pages 73–78.

Guo, J., Che, W., et al. (2014). Revisiting Embedding Features for Simple Semi-supervised Learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120, Doha, Qatar. Association for Computational Linguistics.

Jurafsky, D. y Martin, J. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Klinger, R. y Tomanek, K. (2007). Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology.

Koller, D. y Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Kumar, E. (2008). *Artificial Intelligence*. I.K. International Publishing House Pvt. Limited.

Lafferty, J. D., McCallum, A., et al. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lample, G., Ballesteros, M., et al. (2016). Neural Architectures for Named Entity Recognition. In *In proceedings of NAACL-HLT (NAACL 2016).*, San Diego, US.

Liang, P. (2005). Semi-supervised Learning for Natural Language. Master's thesis, Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology.

Lin, X. (2011). Fine-grained Named Entity Classification in Machine Reading. Master's thesis, University of Oxford.

Manning, C. D. y Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Mikolov, T., Chen, K., et al. (2013a). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., et al. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In Burges, C., Bottou, L., et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miller, S., Guinness, J., et al. (2004). Name Tagging with Word Clusters and Discriminative Training. In Susan Dumais, D. M. y Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 337–342, Boston, Massachusetts, USA. Association for Computational Linguistics.

Murthy, R. y Bhattacharyya, P. (2016). A Deep Learning Solution to Named Entity Recognition. In *COLING 2016, 27th International Conference on Computational Linguistics, April 3-9, 2016, Konya, Turkey*.

Nivre, J. (2000). Logic Programming Tools for Probabilistic Part-of-Speech Tagging. Technical report, School of Mathematics and Systems Engineering.

Passos, A., Kumar, V., et al. (2014). Lexicon Infused Phrase Embeddings for Named Entity Resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan. Association for Computational Linguistics.

Ponce Gallegos, J. C., Torres Soto, A., et al. (2014). *Inteligencia Artificial*. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn).

Poulsen, S. (2005). *Collocations as a language resource. A functional and cognitive study in English phraseology*. PhD thesis, Institute of Language and Communication. University of Southern Denmark.

Ratinov, L. y Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Recasens, M., Màrquez, L., et al. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Recasens, M. T. M. A. M. M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco. European Language Resources Association (ELRA). ACL Anthology Identifier: L08-1222.

Reese, R. (2015). *Natural Language Processing with Java*. Community Experience Distilled. Packt Publishing.

Stern, R., Sagot, B., et al. (2012). A Joint Named Entity Recognition and Entity Linking System. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, HYBRID '12, pages 52–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Szarvas, G. (2008). *Feature Engineering for Domain Independent Named Entity Recognition and Biomedical Text Mining Applications*. PhD thesis, Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and the University of Szeged.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tjong Kim Sang, E. F. y De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 142–147, Stroudsburg, PA, USA. Association for Computational Linguistics.

Turian, J., Ratinov, L., et al. (2010). Word Representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wallach, H. M. (2004). Conditional Random Fields: An Introduction. Technical Report No. MS-CIS-04-21., University of Pennsylvania. Department of Computer & Information Science.

Yang, Y. y Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yang, Z., Salakhutdinov, R., et al. (2016). Multi-Task Cross-Lingual Sequence Tagging from Scratch. *CoRR*, abs/1603.06270.

Yu, M., Zhao, T., et al. (2013). Compound Embedding Features for Semi-supervised Learning. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 563–568.