Intelética 2(3) (2025), 17-27
Revista de Inteligencia Artificial, Ética y Sociedad

INTELETICA

https://inteletica.iberamia.org/

Uso de reconocimiento de entidades nombradas para clasificación de información sobre conflicto armado

Use of named entity recognition for classification of armed conflict information

Ana María Tangarife Patiño [1] ,Cristian David Gutiérrez Céspedes [2]

[1] Escuela Interamericana de Bibliotecología, Universidad de Antioquia
[2] Instituto de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Antioquia

[1] ana.tangarife@udea.edu.co
[2] cdavid.gutierrez@udea.edu.co

"Artículo revisado y recomendada su publicación por el Dr. Pablo Calleja del *Ontology Engineering Group* de la Escuela Técnica Superior de Ingeniería Informática de la Universidad Politécnica de Madrid y por el Dr. Joaquín García Palacios, Catedrático de la Universidad de Salamanca"

Resumen En este artículo se explica la técnica de reconocimiento de entidades nombradas como una forma de extraer información semánticamente de un conjunto de textos sobre conflicto armado colombiano. Se exponen los principales enfoques y tareas de NER y se describen los tipos de entidades nombradas en este dominio para establecer cómo estas están representadas en el discurso y cuál es el nivel de ambigüedad que se presenta en este ámbito especializado. Luego se indican las técnicas, el proceso de anotación, el entrenamiento del corpus, las métricas y resultados, así como los ajustes a los parámetros para la construcción de un modelo de clasificación, según distintas condiciones técnicas, lingüísticas y conceptuales, tanto del corpus como del dominio mismo.
Palabras clave: Clasificación de información, Conflicto armado, Reconocimiento de Entidades Nombradas.

Abstract: This paper explains the named entity recognition technique as a way of semantically extracting information from a set of texts on the Colombian armed conflict. The main approaches and tasks of NER are presented and the types of named entities in this domain are described in order to establish how they are represented in the discourse and what is the level of ambiguity present in this specialized domain. The techniques, the annotation process, the corpus training, the metrics and results, as well as the adjustments to the parameters for the construction of a classification model, according to different technical, linguistic and conceptual conditions, both of the corpus and of the domain itself, are then indicated.
Keywords: Information classification, armed conflict, Named Entity Recognition.

1. Introducción

El estudio sobre el conflicto armado en Colombia, fundamental para entender dinámicas y narrativas sobre el mismo, ha generado gran cantidad de información como informes de investigación, documentación legal, testimonios, registros de prensa, entre otros. Esta información, que es utilizada por investigadores, expertos y comunidad en general suele ser dispuesta en repositorios y bases de datos de instituciones especializadas en el tema, sin embargo, acceder y disponer de estos contenidos tiene grandes dificultades, desde el registro hasta la recuperación de información, que generalmente se da a partir de campos definidos en bases de datos de modelo relacional.

ISSN: 3020-7444

Por otro lado, el reconocimiento de entidades nombradas (NER) hace parte del proceso de extracción de información y se refiere a la detección de menciones de entidades del mundo real a partir de un texto para clasificarlas en categorías predefinidas como lugares, personas, organizaciones, que puedan ser aplicadas posteriormente en la modelación de intereses de usuarios, en sistemas pregunta/respuesta y en sistemas de diálogo [13], comprensión de textos, recuperación de información, resumen automático de textos, traducción automática y construcción de bases de conocimiento [12], desarrollo de *chatbots*, analizadores de contenido u opiniones de consumidores [7] o para la anotación semántica y el poblamiento automático de ontologías [15].

La resolución de entidades tiene grandes desafíos en la extracción como la ambigüedad, la cantidad de datos de entrenamiento, las variaciones propias en un dominio específico, las formas diferentes de una misma entidad [9]. La extracción de entidades pasa también por la identificación de relaciones entre ellas, lo cual es requerido para definir patrones sintácticos, aprendizaje supervisado, extracción de información abierta. Para llevar a cabo esta tarea "un algoritmo NER identifica una entidad nombrada y el tipo al que corresponde, considerando cada palabra en la secuencia y decidiendo a cuál tipo particular pertenece" [10].

Muchas de las tareas del procesamiento de lenguaje natural contemplan el reconocimiento de entidades nombradas para analizar estructuras textuales, extraer términos y sintagmas a partir de estadísticas del corpus, extraer palabras para lematización, clasificar textos basándose en el aprendizaje de modelos semánticos, extraer eventos o reconocer sentidos en oraciones completas.

El interés por estudiar el corpus de este dominio se fundamenta en el reconocimiento de un fenómeno social que existe en Colombia ya por más de seis décadas y que ha implicado la confrontación y conflictos socio-políticos por el enfrentamiento de distintos grupos armados. Este marco de violencia está representado en el discurso cotidiano y en los ámbitos especializados, lo cual ha generado un inmenso volumen de datos.

Este fenómeno, visto como un dominio de conocimiento, incorpora una cantidad considerable de discursos especializados con términos (denominaciones y conceptos) propios de campos involucrados en los estudios sobre el conflicto armado. Además, se incluyen las terminologías propias que han construido organizaciones sociales y comunitarias que trabajan por la defensa de los derechos humanos.

Por lo tanto, se proponen herramientas para clasificación automática que faciliten los procesos de recuperación y acceso a información a partir del sentido de los datos. Una de las ventajas del trabajo con NER es la posibilidad de extraer información de textos no estructurados [11], lo cual se expande a dominios no estandarizados [8]. Además, se propone el uso de NER para el desarrollo ontológico y en la construcción de grafos de conocimiento [18] y [19].

El proceso de reconocimiento de entidades nombradas implica resolver el asunto de las correferencias a partir de la identificación de las menciones propiamente dichas estableciendo su relación con las clases a las que pertenecen. Para esto se explora el uso de modelos de lenguaje que permiten mejor desempeño en las tareas propias del reconocimiento de entidades como la extracción de información, la extracción de relaciones, la normalización de entidades y la validación.

2. Revisión de la literatura

En cuanto al reconocimiento de entidades nombradas con fines de clasificación de información distintos autores [13], [12], [7], [15], [8] y [3] proponen técnicas y modelos en dominios de conocimiento. Es el caso de los trabajos en los campos de la biología molecular, genética, bioquímica y medicina en donde se analizan textos biomédicos para reconocer entidades que puedan llevarse a una ontología [19], o los trabajos de [4], que construyen una ontología sobre patrimonio cultural chino.

El reconocimiento de entidades nombradas requiere el estudio de la terminología [5] y la definición de las clases en el dominio [14], así como la variación terminológica y denominativa [6]. Tanto [9] como [7], abordan el tema de la ambigüedad que es necesario resolver para mejorar los sistemas de reconocimiento y clasificación de entidades.

Una de las ventajas del trabajo con NER es la posibilidad de extraer información de textos no estructurados [11], lo cual se expande a dominios no estandarizados [8]. Además, se propone el uso de NER para el desarrollo ontológico y en la construcción de grafos de conocimiento [18] y [19].

La propuesta de uso de aprendizaje profundo para la extracción de entidades tiene gran interés en la evaluación de los modelos para valorar principalmente la precisión y recuperación de entidades [12].

2.1. Enfoques y tareas para NER

Un sistema para el reconocimiento de entidades nombradas está compuesto por los tipos de entidad que deberán identificarse, la identificación misma, los criterios de anotación, y los límites válidos de identificación de la entidad. Cada uno de estos elementos representa una serie de desafíos tanto semánticos como sintácticos pues se requiere tanto conocer el dominio y la terminología propia de los campos, así como brindar las herramientas conceptuales que serán la base para el reconocimiento.

Los enfoques bajo los cuales se ha abordado el reconocimiento de entidades nombradas son: basados en reglas, aprendizaje no supervisado, aprendizaje supervisado basado en características y aprendizaje profundo. Los enfoques basados en reglas plantean que estas pueden diseñarse a partir de nomenclaturas específicas del dominio y en patrones sintáctico-léxicos que descubren las reglas semánticas y sintácticas. Los sistemas basados en reglas funcionan bien, pero requieren un léxico exhaustivo, pues como afirman [12] "debido a las reglas específicas del dominio y a los diccionarios incompletos, a menudo se observa una alta precisión y una baja recuperación en estos sistemas" (p. 4). Sin embargo, esto requiere la definición de esas reglas lo que demanda un amplio conocimiento en términos del dominio.

Los enfoques de aprendizaje no supervisado parten de la idea de *clusters* que son construidos a partir de recursos léxicos, patrones léxicos y estadísticos que aportarían información sobre grandes corpus para inferir las entidades [12]. El aprendizaje basado en características parte de ejemplos de entrenamiento que hayan sido etiquetados para posteriormente definir algoritmos de aprendizaje que reconozcan patrones para identificar nuevos datos no vistos anteriormente en un corpus más amplio.

Por último, el enfoque que usa el aprendizaje profundo para descubrir información de manera automática resulta de gran interés especialmente en campos en los que no se tienen tantos datos entrenados previamente. Este enfoque utiliza métodos como representación distribuida que pueda darse a nivel de palabras, caracteres o representación híbrida que varía según las propiedades semánticas y sintácticas de las palabras que son identificadas en el texto y la distribución de palabras pre entrenadas.

2.2. Tareas para el reconocimiento de entidades de nombradas

Dentro del reconocimiento de entidades se pueden identificar distintas tareas según necesidades específicas del dominio, distinguiendo al menos tres tipos de tareas: extracción de entidades, extracción de relaciones y extracción de eventos. A estas, se suman otras que están relacionadas con la resolución de correferencias y con la normalización de entidades, de modo que figuras léxicas como la homonimia, la antonimia, la hiperonimia o la hiponimia puedan ser resueltas para garantizar una efectiva representación unívoca del lenguaje.

Resolver las cuestiones de sentido en niveles distintos implica también desglosar las tareas que se demandan para reconocer eficazmente las palabras y agruparlas según las clases a las que pertenecen. En este sentido, [7] propone que un modelo NER debe seguir los siguientes pasos:

- Identificar las frases sustantivas (sintagmas nominales) a partir del análisis sintáctico de dependencias y el etiquetado de cada parte de la oración.
- Clasificar las frases para determinar a qué categorías corresponden, según los diccionarios y otras fuentes conceptuales.
- Desambiguar entidades que puedan estar mal clasificadas, de modo que puedan validarse los resultados.

La detección de entidades implica como tarea también la categorización, según esas entidades se refieran a nombres, localizaciones, eventos, organizaciones. Una categoría o clase puede definirse como un conjunto de elementos que guardan relación entre sí, la palabra por su parte se entiende como una unidad léxica básica para referirse a cosas de la realidad. Se da que "a veces la presencia de palabras individuales es suficientemente informativa para que el algoritmo identifique una clase" [10] (p. 397) y esto en casos en que las palabras posean una carga semántica fuerte y que no lleve a equívocos, como en el caso de nombres propios de organizaciones, por ejemplo, que pueden ubicarse como parte de una tipología de entidad concreta, o en el de palabras de la lengua común cuyo significado puede ser fácilmente atribuible a un concepto u otro, por ejemplo en palabras como *asesinato* o en unidades fraseológicas como *desaparición forzada de personas*.

Las tareas para el reconocimiento de entidades nombradas se describen a continuación:

Extracción de información: Esta extracción implica definir el modo en que se recopila la información de acuerdo con las características de los datos. Siguiendo con [11], "la extracción de información no es más que otra forma de extracción de características de aprendizaje automático a partir de datos de lenguaje natural no estructurados, como la creación de bolsas de palabras o incrustaciones para intentar reducir las casi infinitas

posibilidades de significado del texto en lenguaje natural a un vector que una máquina pueda procesar fácilmente" (p. 345).

Extracción de relaciones: Esto implica la identificación de los patrones en las frases y el sentido que tiene el uso de las palabras y su posición en una frase determinada. Esta relación se da a partir de la identificación del patrón Sujeto - Verbo – Objeto [11] que coincide igualmente con lo que en ontologías se conoce como tripleta de la entidad que recopila la información sobre sujeto o entidad, predicado o atributo y objeto o valor.

Normalización de entidades: Proceso de reconocer los sentidos que tienen las entidades y su validación en el sistema de conocimiento, de modo que haya consistencia en el tratamiento de los datos. La normalización implica tareas en donde se corrijan aspectos tanto formales como semánticos. Tal como lo explicaron [11], "la normalización de las entidades con nombre y la resolución de ambigüedades suele denominarse resolución de correferencias o resolución de anáforas, especialmente en el caso de pronombres u otros «nombres» que dependen del contexto" (p. 357).

Así también, en un clasificador se debe incluir un algoritmo de normalización de modo que cada tipo de entidad tenga un único nombre que se refiera a una misma cosa y que eso sea coherente dentro de la base de conocimientos definida. Por esta razón, es necesario establecer las relaciones de tipo "es-un" para conectar las entidades con las categorías a las que se corresponden. También normalizar elementos como fechas u otros objetos que puedan ser incorporados a una base de conocimiento [11].

Validación: Implica la contrastación del sistema de reconocimiento de entidades versus el reconocimiento y clasificación que haría un humano, quien además fuera experto en el dominio. Es importante definir cuáles son los criterios para evaluar que el reconocimiento de entidades sea adecuado. Para eso, [12] dan cuenta de varios sistemas de evaluación: uno cuando se da la coincidencia exacta entre el sistema NER y el reconocimiento humano para lo cual se definen métricas; la segunda evaluación es de concordancia mínima en la cual se acepta una entidad como correcta siempre que sea reconocida dentro de algunos límites y que no se contradiga con la verdad básica.

Estos procesos de validación implican dos subtareas: la detección de la coincidencia del reconocimiento según los parámetros establecidos y la desambiguación que se haga de las entidades a partir de las características que tienen, sean estos falsos positivos, falsos negativos o verdaderos positivos. El falso positivo se da cuando una entidad es reconocida pero no es verdadera; el falso negativo es cuando la entidad es verdadera pero no es recuperada por el sistema de reconocimiento de entidades. El verdadero positivo se da cuando una entidad es devuelta por el sistema. El modelo NER deberá definir los parámetros de evaluación de coincidencia para determinar la validez de los resultados [17].

2.3. Modelos de lenguaje para construcción de NER

Un modelo de lenguaje es útil porque contiene datos que ya han sido entrenados en los cuales se identifican ya las estructuras sintácticas y gramaticales del lenguaje, que es una tarea básica para poder analizar y reconocer otros elementos relacionados con la semántica.

El uso de estos modelos de lenguaje es aplicable en la resolución de problemas en distintos ámbitos como la educación, la medicina, las finanzas o el servicio al cliente. Para problemas más concretos, se recurre al enfoque de ajuste fino o *fine-tuning*, el cual se usa para entrenar un modelo con datos particulares, es decir que incorporen vocabularios específicos y muestras de texto también acotadas. *Fine-tuning* también es utilizado para ajustar un clasificador en los datos específicos de la tarea para un tratamiento por capas. Entrenar un modelo de lenguaje en un corpus grande y luego ajustarlo en una tarea posterior más pequeña es el enfoque central utilizado en modelos basados en *transformers* y modelos como BERT, GPT, RoBERTa y otros [16].

Fine-tuning es una técnica para entrenar modelos basados en redes neuronales cuando no se dispone de una gran cantidad de datos preentrenados, como sí puede ser el caso de otros trabajos de dominios de conocimiento que usan datos que ya han sido entrenados y anotados anteriormente, y cuyo volumen y validación también es significativo. *Fine-tuning* trabaja a partir de redes neuronales que son arquitecturas que sirven para extraer y aprender directamente desde los datos que son proporcionados a partir del reconocimiento de patrones. Se requiere un conjunto de datos específicos y una muestra de datos anotados que sirvan al modelo como ejemplo de lo que este debería reconocer en su análisis automático.

3. Metodología

Para desarrollar este modelo NER los pasos fueron: adquirir un corpus de textos que contenga entidades como personas, organizaciones, lugares, etc.; limpiar y procesar el corpus de texto haciendo la anotación correspondiente según las clases de entidades definidas; dividir el corpus en conjuntos de entrenamiento, testeo y validación; definir las clases y tipos de entidades que se recuperarán y/o usar modelos de lenguaje incorporados en bibliotecas existentes tipo *Spacy, NLTK, TensorFlow*; evaluar el modelo de prueba y ajustar los parámetros; y, por último, utilizar el modelo para etiquetar datos no entrenados.

3.1. Conformación del corpus (*dataset*)

Para la anotación de un corpus se deben seleccionar muestras que incluyan variedad de textos y que sean representativos del dominio. Este corpus se divide en una parte para el aprendizaje y otra como corpus de prueba, además de ejemplos sin anotar que puedan ser clasificados posteriormente por el algoritmo para validar su eficacia y precisión en el reconocimiento y asignación a una clase.

Como conjunto de entrenamiento se seleccionaron fragmentos de distintas fuentes a partir de textos relevantes en la literatura sobre el conflicto armado colombiano. El *dataset* se constituyó con 658 fragmentos de texto llamados oraciones con una extensión no superior a 50 palabras, resultando en un total de 14485 palabras. Para cada palabra en cada oración se asignó una etiqueta identificando la entidad nombrada a la que pertenece y su posición relativa dentro de dicha entidad.

3.2. Tipos de entidades en el dominio

Una entidad se define como un nombre (propio o común) que sirve para designar algo o alguien. Típicamente son sustantivos o sintagmas nominales que representan un objeto o "cosa" del mundo. Se reconocen dos tipos de entidades: genéricas o específicas. Las primeras estarían representadas en asuntos como personas, lugares, fechas; y las específicas representan conceptos de dominios específicos [12], como para este caso en entidades como *restitución de tierras*, *actores armados*, *dispositivos pedagógicos* o *derecho a la verdad*.

Las entidades nombradas pueden clasificarse en función de los cuatro criterios siguientes: nombre propio, designación rígida, identificación única y ámbito de aplicación [15]. Algunos ejemplos en este dominio son entidades como *Centro Nacional de Memoria Histórica*, *Ruta Pacífica de las Mujeres*, *Jurisdicción Especial para la Paz* para referirse a la clase **ORG** (instituciones u organizaciones); o expresiones que se refieren a la clase **VIO** (hecho de violencia) y que aparecen en entidades como *ejecución extrajudicial*, *desaparición forzada*, *detención arbitraria* o *desplazamiento forzado*. También pueden designar una persona en particular que se identifica con su nombre pero que tiene también una serie de atributos como rol, características sociodemográficas o participación en un hecho. O pueden presentarse algunas entidades que son usadas en otros tantos dominios, como por ejemplo los nombres de lugares o fechas.

En este trabajo fueron identificadas las siguientes clases comunes en las tareas de procesamiento de lenguaje natural, como **GEO** (localización geográfica), **DATE** (fechas y tiempo), **PER** (persona), **ORG** (organización); además se definieron las clases relevantes para describir aspectos específicos del dominio, como son **VIO** (hecho de violencia), **AFE** (afectación) o **ARM** (actor armado). Se describen estas particulares para comprender cuáles son las variaciones sintácticas que se presentan en los textos y el contraste entre las que serían denominaciones conceptuales y propiamente terminológicas y la manera como éstas aparecen en el lenguaje natural o en el corpus. Aquí se presentaron desafíos de orden lingüístico y conceptual que fue necesario ir afinando a medida que se avanzaba en la optimización del algoritmo de clasificación con NER. Uno de esos desafíos es la ambigüedad, que deberá ser resuelta con el concurso de múltiples voces de las distintas especialidades y orillas que componen los discursos de este dominio. La desambiguación requiere la validación por parte de expertos y poder contar con un buen léxico constituido.

3.3. Anotación del corpus

El reconocimiento de entidades nombradas en un corpus de conflictos armados implica el uso de herramientas y técnicas que permitan identificar y clasificar las menciones de personas, lugares y organizaciones que son relevantes para el contexto del conflicto. De acuerdo con [11] "una frase típica puede contener varias entidades con nombre de varios tipos, como entidades geográficas, organizaciones, personas, entidades políticas, tiempos (incluyendo

fechas), artefactos, eventos y fenómenos naturales. Y una oración puede contener varias relaciones entre las entidades nombradas en la oración" (p. 340).

Dentro de las tareas de clasificación para el entrenamiento del modelo, se deben asignar las características y cualidades a las clases que distinguen una de otra. Ello requiere estudiar los significados y usos posibles de las palabras por medio de *Word embeddings*, entendida como una técnica de procesamiento de lenguaje natural para representar las palabras como vectores de números, los cuales capturan la relación semántica entre las palabras. Esta técnica permite además el uso de *Word Sense Disambiguation* para determinar el significado de una palabra específica en un contexto dado.

En este caso, fueron anotados manualmente alrededor de 90.000 *tokens* indicando a cuál de las clases definidas pertenecían cada uno. La anotación se realizó usando el formato BIO, el cual es usado para etiquetar *tokens* en la tarea de segmentación para el reconocimiento de una entidad nombrada de tal modo que se pueda indicar el *token* de inicio y aquellos que son dependientes o hacen parte y que conforman el nombre de entidades e instancias en más de una palabra (sintagmas o frases). En la tabla 1 se describen todas las clases y etiquetas inicialmente definidas para la anotación de los datos.

Clases y labels para anotación

Nombre	Label	Nombre	Label
Sin valor semántico	O	Ley	B-ley I-ley
Persona	**B-per** **I-per**	Fecha	B-date I-date
Organización	**B-org** **I-org**	Luchas y resistencias	B-luc I-luc
Localización	**B-geo** **I-geo**	Memoria	B-mem I-mem
Actor armado	**B-arm** **I-arm**	Atención	B-ate I-ate
Afectación	**B-afe** **I-afe**	Conceptos	B-con I-con
Hecho de violencia	**B-vio** **I-vio**	Paz	B-paz I-paz
Evento	B-eve I-eve	Derechos humanos	B-der I-der

Tabla 1. Clases y etiquetas para anotación

Inicialmente los nombres de las etiquetas tienen una codificación estándar de acuerdo con modelos ya establecidos. Como ya se mencionó, se usaron las etiquetas **PER** (persona), **ORG** (organización), **GEO** (localización geográfica), **DATE** (fecha) y **EVE** (evento) ya estandarizadas tanto para el español como para otros idiomas desde los modelos de lenguaje. Adicionalmente, y para describir el caso concreto, fueron consideradas las etiquetas **ARM** (actor armado), **AFE** (afectación), **LEY** (legislación), **VIO** (hecho de violencia), **LR** (lucha y resistencia), **MEM** (memoria), **ATE** (atención), **CON** (conceptos), **PAZ** (iniciativas de paz) y **DER** (derechos), que corresponden a elementos concretos del dominio. Estas etiquetas fueron utilizadas para la codificación inicial.

Sin embargo, dada la dispersión de los datos se optó por dejar como clases definitivas para la experimentación **ORG**, **PER** y **GEO**, que son comunes en modelos de lenguaje establecidos; y **VIO**, **ARM** y **AFE**, como clases específicas para este dominio, que sirven tanto para la clasificación de información sobre conflicto armado colombiano, pero que puede ser extrapolable a otras tareas de clasificación de información de otros conflictos.

La anotación con estas etiquetas se realiza para el conjunto de los datos que corresponden al *dataset*, es decir, se toman ejemplos del corpus que recogen muestras del uso del lenguaje en textos propios de este dominio de conocimiento.

Para ello se escogen entidades que estén reflejadas en el contexto local del texto, el cual es elegido a partir de criterios de segmentación usando técnicas de *parsing* para establecer la unidad mayor de división en el texto. En este caso se trabajó a partir de oraciones completas dentro de las cuales se etiquetaron las entidades según el formato BIO (*Beginning, Inside, Outside*). Cada *token* debe ser parte de una entidad y debe indicarse a qué parte de la estructura corresponde para lo cual se usa un estándar de etiquetación mediante el cual se establece cuál es el *token* inicial (B), el o los *tokens* que contiene esa entidad (I) y distinguir de los que no tienen ningún valor semántico (O).

Por ejemplo, en la siguiente oración tomada del corpus:

> el 14 de enero de 2004, hombre del Bloque Cacique Nutibara de las Autodefensas Unidas de Colombia, en alianza con bandas conocidas como El Hueco y La 38, al parecer, lideradas por un grupo de reinsertados, asumieron el control territorial y social del barrio Popular Uno de la ciudad de Medellín, ordenando el desplazamiento forzado de varios grupos familiares,

se identifican las entidades: *Bloque Cacique Nutibara*/**ARM**; *Autodefensas Unidas de Colombia*/**ARM**; *El Hueco*/**ARM**; *La 38*/**ARM**; *Popular Uno*/**GEO**; *Medellín*/**GEO**; *Desplazamiento forzado*/**AFE**.

3.4. Modelo de clasificación con NER

El proceso de reconocimiento de entidades nombradas implica resolver el asunto de las correferencias identificando las menciones propiamente dichas y estableciendo a cuáles clases pertenecen.

La unidad básica de procesamiento en tareas de procesamiento de lenguaje natural es el *token*, sin embargo en la mayoría de los casos una entidad está constituida por dos o más *tokens* que la representan íntegramente un concepto o entidad (sintagmas). Si entendemos las entidades como una manera de representar conceptualmente objetos de la realidad, esto se complejiza porque esa representación conlleva al uso de varios *tokens* que corresponden a una estructura sintáctica determinada, pero que un proceso automatizado tendrá que identificar para poder validar cuándo una cadena de caracteres que conforman un conjunto de *tokens* pueden considerarse efectivamente entidades.

Para resolver esto se entrena un modelo para la identificación de las entidades en el que se puedan juntar, por un lado, la tarea de reconocer entidades según modelos de lenguaje establecidos y, por otro, la tarea de vincular terminología propia del dominio, que parte de una conceptualización que se ha sistematizado a partir de las fuentes terminológicas y conceptuales sobre el conflicto armado.

A partir del uso de modelos de lenguaje se incluye el reconocimiento de entidades nombradas de carácter general. Con la incorporación de otros ejemplos con clases del dominio particular, se busca una cobertura más amplia para identificar entidades propias de este ámbito. La vinculación de una terminología propia se realiza a partir de un proceso de etiquetación de ejemplos que nutran el modelo y que sirvan como entrenamiento para el aprendizaje.

La construcción de un modelo de reconocimiento de entidades nombradas debe incluir ejemplos anotados correctamente. Los modelos no aprenden todo, es un proceso de constante revisión y ajuste, por lo cual los esquemas de *labels* o etiquetas deben ser consistentes y no demasiado específicos, pues esto tendería a atomizar tanto la información que luego los resultados sean inexactos y el modelo ineficiente.

El entrenamiento manual implica, por un lado, la conceptualización, es decir establecer qué clases y qué tipos de entidades serán reconocidas; y por otro lado la anotación que implica la segmentación dentro del conjunto de textos para distinguir los *tokens* que son entidades y los que no lo son, y a cuáles clases pertenecen estas entidades. Luego, se usa esta estructura para anotar y para validar el funcionamiento del modelo.

El entrenamiento también debe contemplar otros *tokens* que no son entidades para poder distinguir una cosa de otra y aprender sobre las características que tienen las palabras cuando corresponden a entidades y cuando no. De ese modo el modelo estará preparado para reconocer nuevas entidades que no se han indicado previamente en contextos similares, lo que es justamente una de las virtudes de la utilización del aprendizaje de máquina y el uso de redes neuronales.

Para montar el modelo, se hace una selección de los datos para construir dos archivos: uno para el entrenamiento y otro para testeo. El entrenamiento está constituido por casi 90.000 *tokens* anotados manualmente en los cuales se encuentran 3.760 entidades identificadas. El testeo está conformado por más de 29.000 *tokens* anotados en los cuales se encuentran 1.162 entidades anotadas.

Estos datos se crean a partir del método basado en reglas para generar un conjunto de datos de entrenamiento básico que sirve para definir patrones, que pueden ser cadena y *token*. Cadena cuando son *tokens* que conforman una frase o sintagma y *token* cuando son una sola palabra. Luego, estos datos se llevan a un formato que pueda procesarse dentro del algoritmo y se construye el sistema de clasificación a partir de las etiquetas establecidas. Después se entrena el modelo, tarea que se realiza de acuerdo con los hiperparámetros definidos y se obtienen los resultados.

Definir este modelo implica responder sobre cuestiones de orden teórico y práctico: una es sobre cómo se reconoce una estructura sintáctica para hacer anotación semiautomática, cómo se pueden integrar otros corpus, lexicones y terminologías en el procesamiento dado que muchas de las entidades es posible que puedan ser identificadas y reconocidas por otros modelos o lenguajes o que incluso existan ya en otros recursos. Esto implica, entre otras cuestiones, entrenar la máquina para identificar los elementos que necesitamos que entienda y darle detallada instrucción sobre los pasos a seguir si encuentra una u otra entidad.

Todas estas cuestiones son tenidas en cuenta según la forma del corpus y los elementos que se quieren observar y cuáles son las categorías a partir de las cuales se establece la organización y representación de los datos. Por tanto, todo depende de muchos factores y requiere una revisión cuidadosa de aspectos conceptuales del dominio y lingüísticos según la naturaleza misma de los textos.

De otro lado, *FLAIR* es un paquete de *Python* basado en *Pytorch* y desarrollado en la *Humboldt University of Berlin*, con el fin de facilitar el desarrollo e implementación de experimentos relacionados con tareas de Procesamiento de Lenguaje Natural (NLP) *como Named Entity Recognition* (NER), análisis de sentimientos o *part-of-speech tagging* (PoS) [1]. *FLAIR* presenta tres facilidades a la hora de construir modelos basados en datos:

- La construcción de un *dataset* dividido en las secciones entrenamiento, validación y prueba de manera automática mediante el objeto Corpus.
- La combinación de múltiples *embeddings* para la codificación del texto en vectores con el objeto *StackedEmbeddings.*

En este trabajo se configuró un modelo del tipo *SequenceTagger* con una capa oculta de 256, una pila de *embeddigns* conformada por GloVe y dos *embeddings* propios de FLAIR para el español: es-*forward* y es-*backward*. El *tag dictionary* quedó distribuido de la siguiente manera: **PER** (*seen* 164 *times*), **GEO** (*seen* 116 *times*), **VIO** (*seen* 113 *times*), **ORG** (*seen* 102 *times*), **ARM** (*seen* 65 *times*), **PAZ** (*seen* 42 *times*), **DATE** (*seen* 40 *times*), **AFE** (*seen* 33 *times*). Para el entrenamiento, se configuró un *TrainerModel* con una tasa de aprendizaje de 0.05, un *batch size* de 2 y un número máximo de épocas de 100.

3.5. Evaluación

El problema de las entidades nombradas puede entenderse como un problema de clasificación donde las etiquetas de las entidades son las clases a las que cada palabra en el texto puede pertenecer. Una forma usual de representar los resultados de la clasificación es un arreglo bidimensional llamado matriz de confusión, en donde la entrada representa la cantidad de textos pertenecientes a la clase *i* que fueron clasificados por el modelo como pertenecientes a la clase *j*. La figura 1 muestra la matriz de confusión para dos clases, así como la nomenclatura usual para cada elemento de esta.

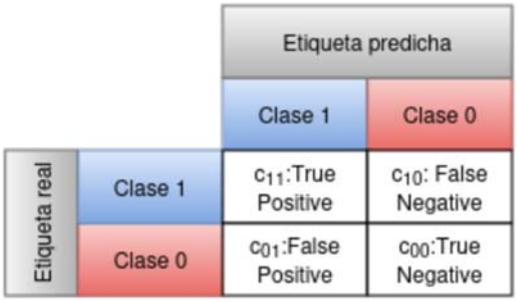

Figura 1: Matriz de confusión para un problema de clasificación binario.

En los problemas de clasificación, tres métricas son las más usuales: *precision*, *recall* y *F1-score* [2].

Precision: es la proporción de verdaderos positivos respecto a todos los clasificados como positivos:

Etiqueta	Precision	Recall	f1-score
PER			
GEO			
VIO			
ORG			
ARM			
PAZ			
DATE			
AFE			

Tabla 2. Resultados en el entrenamiento

Recall: o sensibilidad es una medida de cuántas de las predicciones hechas para una clase, de verdad pertenecen a esa clase. Está dado por:

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: es la media armónica de *precision* y *recall* de un clasificador. Resulta de mayor relevancia en el caso de *datasets* no balanceados. El *F1 score* se calcula como:

$$F_1 = 2 \cdot \frac{\mathrm{Pr}\,e\,cision \cdot Recall}{\mathrm{Pr}\,e\,cision + Recall}.$$

4. Resultados y discusión

La evaluación de un modelo de reconocimiento de entidades nombradas es una tarea crítica en el procesamiento de lenguaje natural puesto que implica la identificación y extracción en un texto distinguiendo las que son personas, lugares, organizaciones o fechas, que están predefinidas ya en modelos de lenguaje; pero también reconociendo otras que en este ejercicio particular fueron anotadas para distinguir elementos que en el dominio son importantes como hechos de violencia, afectaciones o actores armados, por ejemplo.

Para evaluar el rendimiento de un modelo de NER es común utilizar medidas de evaluación como la precisión, el *recall* y la puntuación F1. La precisión mide la fracción de las entidades identificadas por el modelo que son correctas, mientras que el *recall* mide la fracción de las entidades presentes en el texto que el modelo fue capaz de identificar. La puntuación F1 combina la precisión y el *recall* en una sola medida que proporciona una evaluación más completa del rendimiento del modelo.

En cuanto a los datos de entrenamiento, es importante tener una muestra diversa y representativa de textos que cubran una variedad del dominio con sus respectivos contextos y que incluyan una amplia gama de entidades con nombre. Para etiquetar los datos de entrenamiento, se pueden utilizar herramientas de anotación manual o semiautomática, y se deben seguir las directrices y estándares de etiquetado establecidos para NER. En este caso se efectúa una anotación manual y se hace una selección de textos más cuidada.

Es importante destacar que la calidad de los datos de entrenamiento es esencial para el rendimiento de un modelo de NER. Si los datos de entrenamiento no son adecuados, el modelo puede tener dificultades para generalizar a textos nuevos y desconocidos, lo que puede resultar en una precisión y un *recall* bajos. Por lo tanto, es importante asegurarse de que los datos de entrenamiento sean lo más representativos y de alta calidad posible.

Una de las dificultades al construir este modelo es que las métricas no eran óptimas pues se presentaba una gran dispersión de los datos, por lo que fue necesario acotar y dejar algunas de las clases y no todas las que se habían definido inicialmente y sobre las que vale la pena explorar más adelante mediante la anotación de más recursos.

A continuación se presentan algunas de las consideraciones que se tuvieron para hacer los ajustes del modelo, en aspectos que van desde el proceso mismo de anotación, pasando por asuntos más conceptuales, hasta las mismas métricas y validación.

Cuando las entidades son más denominativas y representan lo que se entendería como un término propiamente dicho, el modelo puede predecir mejor; pero cuando las entidades corresponden a palabras del léxico común es más difícil identificar que estas pertenecen a una clase determinada. Esto demanda que haya una buena cantidad de ejemplos anotados que permitan distinguir y reconocer los patrones que faciliten la predicción.

En el proceso de anotación se usan los ejemplos tomados del corpus, pero por otro lado se crean también unos diccionarios en donde se listan entidades ya identificadas y que se corresponden a una de las clases definidas como hechos de violencia o actores armados. Se encuentran por ejemplo, unidades como *abuso de menores*, *abusos contra pueblos indígenas*, que podrían considerarse conceptos y que fueron etiquetados usando los mismos *labels* de las clases definidas. Contar con un diccionario es interesante porque permite ampliar las posibilidades del modelo a partir de recursos léxicos constituidos. Por ejemplo, puede ser que entre los ejemplos anotados no aparezca un término como *allanamiento masivo*, pero que en los diccionarios sí aparezca dado que se considera un término.

La anotación requiere que haya consistencia, sin embargo, esto no es siempre sencillo dada la ambigüedad semántica o el significado relativo que pueden adquirir las palabras según el contexto de aparición. Por ejemplo, un *token* solo como *desplazamiento*, puede referirse tanto a la clase **VIO** (hecho de violencia) como **AFE** (afectación).

5. Conclusiones

En cuanto al Reconocimiento de Entidades Nombradas (NER), se encuentra que construir herramientas en dominios particulares demanda la existencia de buenos léxicos para la extracción de términos y estos léxicos podrían ser, bien las ontologías o bien los corpus anotados, ejemplos en donde equipos y personas hacen el trabajo de clasificación y anotación para reconocer las entidades y que estos sean validados para garantizar que las herramientas recogen un saber consensuado.

La tarea de reconocer entidades nombradas se ve enfrentada al fenómeno de la ambigüedad semántica y, por tanto, definir criterios de desambiguación es fundamental. Incluso para ámbitos de la ciencia, en donde se pretende una univocidad y estandarización, se presentan problemas de ambigüedad descritos en áreas como las ciencias biomédicas [19] o la química [20]. Una de las tareas planteadas para desambiguar sentidos en el texto es identificar dónde está posicionado un concepto, lematizar con reglas y a través de *word embeddings* reconocer cuáles de las palabras que se encuentran permitirán ayudar a predecir el sentido que esa palabra tiene y la localización en el gran mapa conceptual de su dominio.

En esta tarea se demuestra que usar modelos de lenguaje es de gran utilidad en ámbitos generales, pues ellos incorporan cierto conocimiento del mundo ya muy estandarizado como en el caso de los nombres de organizaciones, lugares o personas. Sin embargo, cuando se aplican al análisis de corpus de dominios específicos, hay una gran cantidad de conocimiento que no se encuentra aún representado. Y por tanto la inclusión de nuevas entidades de conocimientos específicos a modelos más amplios permitirán la explotación mayor de estas herramientas para favorecer ya no solo la representación sino la extracción, la clasificación y el análisis.

En la exploración de herramientas y técnicas para la clasificación a partir de NER se encontraron muchas bibliotecas, librerías y recursos, muchas de las cuales son de código abierto y adaptables a necesidades particulares, lo cual es de gran utilidad, pero se reconoce también una curva de aprendizaje importante para interactuar con ellas. Otra opción es el uso de modelos de lenguaje, que también se exploraron. Se define el uso de la técnica *fine-tuning* para construir una herramienta particular y entrenar el modelo a partir de las clases y la base conceptual definida para la modelación del dominio.

Agradecimientos. Este artículo es derivado de la tesis “Modelo semántico y computacional para análisis del conflicto armado en Colombia” del Doctorado en Traducción y Ciencias del Lenguaje de la Universitat Pompeu Fabra, cuya realización tuvo el apoyo de la Universidad de Antioquia y la Fundación Carolina.

Referencias

[1] Alan Akbik et al. (2019). “FLAIR: An easy-to-use framework for state-of- the-art NLP”. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, 54–59.

[2] Bruce P.; Bruce, A. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts. O’Reilly Media, 2017. ISBN: 9781491952917. URL: https://books.google.com.co/books?id=ldPTDgAAQBAJ

[3] Das, S., Katiyar, A., Passonneau, R., Zhang, R. (2021). CONTAINER: Few-Shot Named Entity Recognition via Contrastive Learning. *arXiv:2109.07589.* https://doi.org/10.48550/arXiv.2109.07589

[4] Dou, J., Qina, J., Jina, Z., Lia, Z. (2018). Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. Journal of Visual Languages and Computing, (48), 19–28. https://doi.org/10.1016/j.jvlc.2018.06.005

[5] Estopà, R. (1999). *Extracció de terminologia: elements per a la construcció d’unSEACUSE(Sistema d’Extracció Automàtica de Candisats a Unitats de Significació Especialitzada).* Universitat Pompeu Fabra.

[6] Freixa, J. (2005). Variación terminológica: ¿Por qué y para qué?. *Meta: journal des traducteurs / Meta: Translators' Journal*, 50(4). https://doi.org/10.7202/019917ar

[7] Goyal, C. (2021) Part 10: Step by Step Guide to Master NLP – Named Entity. *Analytics vidhya*. https://www.analyticsvidhya.com/blog/2021/06/part-10-step-by-step-guide-to-master-nlp-named-entity-recognition/

[8] Gupta, N., Singh, S., Roth, D. (2017). Entity Linking via Joint Encoding of Types, Descriptions, and Context. En: Palmer, M., Hwa, R., Riedel, S. (Ed.) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (pp. 2681–2690). https://aclanthology.org/D17-1284/

[9] Kejriwal, M., Knoblock, C., Szekely, P. (2021). *Knowledge Graphs: Fundamentals, Techniques, and Applications.* The MIT Press.

[10] Kochmar, E. (2022). *Getting Started with Natural Language Processing*. Manning.

[11] Lane, H., Howard, C., Hapke, H. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning.

[12] Li, J., Sun, A., Han, J., Li, C. (2020). A Survey on Deep Learning for Named Entity Recognition. *arXiv:1812.09449.* https://doi.org/10.48550/arXiv.1812.09449

[13] Liang, C., Yu, Y., Jiang, H., Er, S., Wang, R., Zhao, T., Zhang, C. (2020). BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. *arXiv:2006.15509.* https://doi.org/10.48550/arXiv.2006.15509

[14] Marneffe, M.; Manning, C., Nivre, J., Zeman, D. (2021). Universal dependencies. *Computational linguistics*, 47(2), 255-308. https://doi.org/10.1162/coli_a_00402

[15] Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J. (2013). Named Entity Recognition: Fallacies, Challenges and Opportunities. *Computer Standards & Interfaces,* 35(5), 482-489. https://doi.org/10.1016/j.csi.2012.09.004

[16] Raschka, S. (2023). Understanding Large Language Models: A Cross-Section of the Most Relevant Literature To Get Up to Speed. *Ahead of AI.* https://magazine.sebastianraschka.com/p/understanding-large-language-models

[17] Ruder, S. (2022). *Entity Linking: Task.* http://nlpprogress.com/english/entity_linking.html

[18] Santoso, J., Setiawana, E., Purwantob, C., Yuniarnoc, E., Hariadic, M., Purnomo, M. (2021). Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short-term memory. *Expert Systems With Applications* (176). https://doi.org/10.1016/j.eswa.2021.114856

[19] Wang, K., Stevens, R., Alachram, H., Li, Y., Soldatova, L., King, R., Ananiadou, S., Schoene, A., Li, M., Christopoulou, F., Ambite, J., Matthew, J., Garg, S., Hermjakob, U., Marcu, D., Sheng, E., Beißbarth, T., Wingender, E., Galstyan, A., Gao, X., Chambers, B., Pan, W., Khomtchouk, B., Evans, J. (2021). NERO: a biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding. *Systems Biology and Applications,* 7(1). 1-8. https://doi.org/10.1038/s41540-021-00200-x.

[20] Wang, X., Hu, V., Song, X., Garg, S., Xiao, J., Han, J. (2021). ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-Guided Distant Supervision. En: Moens, M., Huang, X., Specia, L., Wen-tau Yih, S. (Ed.). *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (5227–5240). Association for Computational Linguistics. https://aclanthology.org/2021.emnlp-main.424/