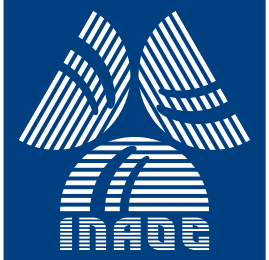

Clasificación de Entidades Nombradas utilizando Información Global

por

Carolina Rocío Sánchez Pérez

Tesis sometida como requisito parcial para obtener el grado de

MAESTRO EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS DE LA COMPUTACIÓN

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Noviembre 2008

Tonantzintla, Puebla

Supervisada por:

Dr. Manuel Montes y Gómez, INAOE

Dr. Aurelio López López, INAOE

A mi mamá

Que me ha dado la vida y su amor en todo momento.

A Rigoberto

Que ha sido mi aliento y fuerza para seguir siempre adelante.

Índice general

Agradecimientos VII

Resumen IX

Abstract XI

1. Introducción 1

1.1. Problemática 4

1.1.1. Objetivos 8

1.2. Organización de la Tesis 9

2. Conceptos Básicos 11

2.1. Aprendizaje Automático 11

2.1.1. Clasificación 12

2.2. Algoritmos de Aprendizaje 13

2.2.1. Árboles de Decisión 13

2.2.1.1. Representación de Árboles de Decisión 14

2.2.1.2. Algoritmo ID3 15

2.2.2. Evaluación de la Clasificación 16

2.3. Reconocimiento de Entidades Nombradas 17

2.3.1. Métricas de Evaluación en REN 20

3. Trabajo Relacionado **23**

3.1. Aproximaciones para CEN 23

3.1.1. Aproximaciones “Manuales” 24

3.1.2. Aproximaciones con Aprendizaje Automático 26

3.1.3. Enfoques de Clasificación 28

3.1.3.1. Clasificación en un Contexto Local 28

3.1.3.2. Clasificación en un Contexto Global. 29

4. Arquitectura del Método Propuesto **35**

4.1. Arquitectura General 35

5. Vinculación de Entidades Nombradas **41**

5.1. Método de Vinculación 41

5.1.1. Medidas de Similitud 43

5.2. Evaluación de Vinculación de EN 47

5.2.1. La Métrica de Evaluación B-CUBED 47

5.2.2. Resultados de Vinculación de Entidades 51

5.2.3. Proceso de Validación de Resultados 53

6. Clasificación Utilizando Información Global **59**

6.1. Enfoques Basados en Voto 60

6.1.1. Voto Simple 61

6.1.2. Voto Ponderado 61

6.1.3. Resultados de los Enfoques Basados en Voto 62

6.1.3.1. Conjunto de Datos 62

6.1.3.2. Resultados 66

6.1.4. Análisis y Variantes de Enfoques Basados en Voto 68

6.1.5. Análisis de Distribución de Clases en Cadenas 75

6.2. Enfoques basados en Árboles de Decisión 81

6.2.1. La Información Global como Atributos 82

6.2.2. Resultados de Árboles de Decisión 88
6.3. Voto entre Documentos 90
6.4. Análisis de Valores Ideales 93

7. Conclusiones y Trabajo Futuro **97**
7.1. Resumen 97
7.2. Conclusiones 98
7.3. Trabajo Futuro 100

Bibliografía **111**

Agradecimientos

A mis asesores, Dr. Manuel Montes y Gómez y Dr. Aurelio López López por su guía, apoyo, consejos y comentarios en la dirección de este trabajo.

A mi comité de tesis, Dr. José Francisco Martínez Trinidad, Dr. Saul Pomares Hernández y Dr. Luis Villaseñor Pineda por sus consejos para el enriquecimiento de este trabajo de tesis.

A mi familia quienes con su ayuda, apoyo y comprensión me alentaron a lograr una más de las metas en mi vida.

A Rigoberto, gracias por tu comprensión y confianza, amor y amistad, por creer en mi, aún en mis momentos difíciles.

Durante mi estadía en el INAOE conocí grandes personas e hice grandes amigos, Roberto, Hugo, Pato, Artemio y Nadia gracias por la compañía y risas en estos dos años.

Mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca No. 204444.

Resumen

El reconocer y clasificar nombres de personas, lugares, organizaciones o cantidades, es un paso importante, y en la mayoría de los casos indispensable en distintas aplicaciones del Procesamiento del Lenguaje Natural (PLN), este proceso es el objetivo del Reconocimiento de Entidades Nombradas (REN).

El proceso de reconocimiento de entidades nombradas se divide generalmente en dos pasos: la delimitación de entidades nombradas (ENs) y su posterior clasificación, en este trabajo de investigación nos enfocaremos en esta última. En la mayoría de las aplicaciones, la clasificación se resuelve basándose en un contexto local, estos métodos no aprovechan la información global que brindan las distintas menciones de una EN para alcanzar un mejor desempeño en la tarea de clasificación. Este método de utilizar la información global de un documento es el que se propone en este trabajo de investigación, utilizando la información del contexto de las referencias de una EN y no sólo la referencia en si. En este trabajo se presenta un enfoque diferente al tradicional mediante la integración de dos pasos, la vinculación de ENs y el refinamiento de la clasificación inicial. El propósito de esto es utilizar la mayor cantidad de información posible, disponible en las distintas menciones de las ENs y apoyar una segunda clasificación.

El proceso de vinculación se basa en medir la similitud entre ENs utilizando distintas medidas para determinar un vinculo entre ENs. En cuanto al refinamiento de clasificación se utilizan enfoques basados en voto y basados en

árboles de decisión, los primeros basados en la teoría de ensambles al combinar clasificaciones obtenidas en un primer paso; los segundos basados en árboles de decisión para integrar la información de las distintas menciones y las cadenas como atributos. Los resultados experimentales con los distintos métodos no presentan una mejora sustancial con respecto a la clasificación inicial obtenida, sin embargo, se considera que en un dominio más específico los métodos propuestos, como en el caso de los métodos de voto, tendrán una influencia positiva con respecto a la clasificación inicial; además de que este trabajo representa una pauta para analizar el impacto de la información global en distintos dominios y escenarios para el idioma español.

Abstract

The recognize and classify the names of people, locations, organizations, or quantities, is an important step, and in most cases, indispensable in different applications of Natural Language Processing (NLP), this process is the objective of the Named Entities Recognition (NER).

The process of recognition of named entities is generally divided into two steps: the delimitation of named entities (NE) and their subsequent classification, in this research work we will focus on the latter. In most of the applications, classification is solved based in a local context, these methods don´t exploit the global information provided by various references in order to achieve a better performance in the classification task. This method of using global information information of a document is what is proposed in this research, using information from the context of the NE references and not only the reference by itself. This paper presents a different approach to the traditional, by integrating two steps, the NEs linking and refining of the initial classification. The purpose of this approach is to use as much information as possible, available in the different references to the ENs and support a second classification.

The process of linking its based in measure the similarity between NEs using different measures to determine a link between ENS. As for the refinement of classification, vote-based approaches and decision trees approaches are used, the first based on the theory of assemblages by combining classifications obtained in a first step, the latter based on decision trees to integrate in-

formation from different references and chains like attributes. The experimental results with different methods do not present a substantial improvement with regard to the initial classification obtained, however, represent a base to analyze these approaches with different domains and scenarios, where it is believed could make more substantial improvements.

Capítulo 1

Introducción

Desde la aparición de distintas tecnologías, como la Internet, la cantidad de información disponible en distintos formatos y fuentes ha crecido a pasos agigantados. El tamaño de las colecciones almacenadas dificulta su manejo y organización, así como la posibilidad de encontrar información específica, ya sea, en un solo documento o en un conjunto de documentos. Además, la información puede no tener una estructura definida, como es el caso de la información textual, donde se encuentran principalmente secuencias de palabras. Dadas estas dificultades el desarrollo de herramientas que permitan administrar y permitir la búsqueda de ciertos elementos en un documento se ha vuelto necesario y de suma importancia.

Con el paso de los años se han hecho esfuerzos en facilitar la comunicación hombre-computadora por medio del lenguaje humano (o lenguaje natural). El Procesamiento del Lenguaje Natural (PLN) nació como una disciplina encargada de desarrollar técnicas computacionales que posibiliten dicha comunicación, donde estas técnicas deben ser capaces de procesar el lenguaje humano, ya sea hablado o escrito, utilizando algún tipo de conocimiento lingüístico. Y es el manejo de este tipo de conocimiento lo que distingue a las aplicaciones del procesamiento del lenguaje de los sistemas de procesamiento

de datos [16].

Con el paso de los años se ha logrado un gran avance en distintas áreas del PLN, se han desarrollado y comercializado aplicaciones en áreas como el Reconocimiento del Habla, Recuperación y Extracción de Información, Generación Automática de Resúmenes, Traducción Automática, entre otras.

En estas áreas, un problema común es obtener información relevante relacionada con nombres de personas, lugares u organizaciones, por lo cual se vuelve importante el poder extraer y distinguir este tipo de elementos de todo el conjunto de palabras que componen a un documento. Aún cuando algunos elementos son relativamente fáciles de identificar, mediante el uso de patrones (por ejemplo fechas o datos numéricos) existen otros elementos, como personas, lugares u organizaciones, que presentan otras dificultades para ser identificados como pertenecientes a un tipo específico. El extraer y distinguir este tipo de elementos es el objetivo del Reconocimiento de Entidades Nombradas (REN).

Una área donde es vital el REN es el área de la Extracción de Información (EI), dentro de esta disciplina los sistemas de extracción de información realizan la tarea de buscar información muy concreta en colecciones de documentos, extrayendo y organizando la información relevante. También en el área de Búsqueda de Respuestas y en Recuperación de Información el reconocer una Entidad Nombrada (EN) sirve para reducir el espacio de búsqueda [32, 25], en la generación de resúmenes la identificación de sentencias relevantes generalmente se relaciona con identificar ENs que están inmersas en estas sentencias [13, 11]; en el caso de Traducción Automática es importante reconocer ENs porque generalmente estas permanecen sin traducción y etiquetarlas para que permanezcan intactas permite mantener coherencia en los textos [2]; estos son sólo algunos ejemplos de la aplicación del REN como parte fundamental en distintas áreas del PLN.

En los últimos años se ha trabajado ampliamente en el desarrollo de sis-

temas de reconocimiento de Entidades Nombradas que van encaminados a mejorar el desempeño de clasificadores utilizando técnicas de aprendizaje automático, mejorar sistemas basados en recursos construidos a mano, aplicaciones orientadas a la extracción de información, entre otros. Todos dirigidos a proponer distintas soluciones al problema. El interés ha sido motivado por el impacto que la solución puede tener en muchas aplicaciones y el potencial comercial de estas. A continuación se listan algunos ejemplos:

- Máquinas de búsqueda más exactas. Encontrar referencias a Lázaro Cárdenas, Michoacán, sin perder tiempo en páginas con información acerca del Presidente Lázaro Cárdenas.
- Antes de leer un artículo o un documento de interés un usuario podría obtener una lista de las personas, lugares y compañías mencionadas en el documento.
- Indexado automático de libros. Para muchos libros y documentos la mayor parte de los términos más informativos respecto a su contenido son ENs.
- Revistas y periódicos en línea podrían resaltar nombres de personas mencionadas, fechas importantes, lugares donde ocurren los hechos, etc.
- El etiquetado de ENs puede servir como un paso de pre-procesamiento para simplificar tareas como por ejemplo traducción por computadora.
- Además el REN es un componente esencial de tareas de extracción de información más complejas, por ejemplo, eventos donde se relacionan nombres de personas, organizaciones, lugares y fechas.

El término *“Entidad Nombrada"* (EN) se utiliza ampliamente en las áreas

mencionadas de PLN, la definición empleada en estos campos se da a continuación.

Una *entidad nombrada* es una palabra o secuencias de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad [27].

Tomemos por ejemplo las siguientes sentencias, donde se tiene resaltada la EN, **Benito Juárez**:

- ***Benito Juárez*** *fue el presidente de la República Mexicana de 1858 a 1872.*
- *El distrito* ***Benito Juárez*** *revocó los permisos de construcción y uso de tierra.*
- *El partido entre Lobos y León se jugará el 24 de mayo en el estadio* ***Benito Juárez.***

Para una persona es fácil identificar que en la primera oración se hace referencia a una persona, en la segunda a una organización política y en la tercera a un lugar, pero esto da lugar al problema de la clasificación de ENs (CEN) que no necesariamente es del mismo tipo. Estos casos ambiguos son sólo uno de los puntos que hacen más difícil el obtener mejores niveles de desempeño en la tarea. Puntos importantes a tomar en cuentan son: ¿Cómo es que se puede atacar este problema de clasificación? ¿Qué características de la sentencia y de su contexto pueden ayudar a reconocer y clasificar las ENs?

1.1. Problemática

El problema de Reconocimiento de Entidades Nombradas generalmente se resuelve como un proceso en dos fases: la delimitación de ENs y su clasificación.

La delimitación de ENs involucra definir el inicio y fin de una EN, es decir, delimitar qué palabras pertenecen a la EN. A su vez, la clasificación consiste en decidir de qué tipo es (lugar, organización, persona u otro tipo) la palabra o palabras delimitadas.

La Clasificación de Entidades Nombradas (CEN) es una tarea de lingüística computacional que busca clasificar cada palabra de un documento en una de las categorías siguientes: persona, lugar, organización, miscelánea o ninguna de las anteriores (por mencionar algunas clases). Como ya se mencionó, la CEN tiene aplicación en distintas tareas del PLN, se han construido distintos sistemas y se han hecho avances en esta área, sin embargo existe un gran número de casos ambiguos que hacen difícil alcanzar mayores niveles de desempeño en dichos sistemas.

Es precisamente la segunda fase del REN, la clasificación, la que presenta mayores dificultades y en la que se centrará este trabajo, viendo la delimitación de ENs como una caja negra.

Trabajo previo se ha realizado en la clasificación de entidades utilizando la información local que puede brindar la palabra que se va a clasificar; como por ejemplo, examinar características de las 2 palabras vecinas de la entidad y asignarles etiquetas, para entonces tomarlas como características locales de la entidad a clasificar [23, 28]. El término información local se refiere a las palabras que componen a la EN a clasificar, así como los elementos que la rodean, generalmente los vecinos a la derecha y a la izquierda de la EN. También distintos atributos de las palabras a clasificar y los atributos de sus vecinos forman parte de la información local.

Supongamos que se tiene la oración:

El Recinto de Homenaje a don ***Benito Juárez*** *fue inaugurado oficialmente el 18 de julio de 1957*.

Benito Juárez es una EN y su información local la compondrían los atributos de las palabras *Benito* y *Juárez* (por ejemplo características ortográficas

o las palabras en si) pero también se pueden integrar características de su contexto, como por ejemplo la palabra *don* que antecede a *Benito* y la palabra *fue* que le sigue a *Juárez*. Ya sea que se integren las palabras u otras características (como su raíz), esta información sería parte de la información local de cada EN.

En la clasificación, generalmente las palabras que rodean a la entidad que se va a clasificar, es decir su información contextual, presenta ciertas características que permitirán distinguirla de entre un conjunto de clases que será definido. De esta forma, aún cuando una entidad nombrada no se haya visto anteriormente, su contexto es el que permite realizar la clasificación, en el ejemplo anterior la palabra *don* puede ser una palabra "disparadora" que indique la existencia de una EN de tipo Persona.

Aún con esto existirán casos donde puede ser difícil predecir o clasificar una entidad nombrada que no ha sido vista debido a un contexto local pobre, pero en muchos casos se tienen distintas referencias de la EN (coreferencias a una EN). Las referencias serán información diferida en un documento (contexto global) que puede ser útil para reforzar el contexto local en el que se encuentra la EN y entonces hacer una correcta clasificación.

Consideremos ahora el siguiente fragmento para ejemplificar algunas de las dificultades en la clasificación:

El Recinto de Homenaje a don ***Benito Juárez*** *fue inaugurado oficialmente el 18 de julio de 1957. Está instalado en el mismo lugar en que el presidente* ***Juárez*** *vivió y murió, en el ala norte de Palacio Nacional, domicilio marcado, en los tiempos del Benemérito, con el número 1 de la calle de Moneda. En 1861, después del triunfo liberal de la Guerra de Reforma,* ***Juárez*** *instaló su gobierno en la capital de la República. La familia Juárez-Maza, al fin reunida, vivió en aquel lugar los momentos de mayor intimidad doméstica, disfrutando, por fin de la paz que* ***Juárez*** *había logrado para toda la nación. El 18 de julio de 1872,* ***Benito Juárez*** *fallecía en la que había sido su habitación conyugal, en la casa*

de su familia que se convertiría, años después, en el recinto de homenaje a su memoria.

Podemos ver que en algunas partes se hace referencia a una misma entidad nombrada *Benito Juárez* pero con un conjunto de referencias que no necesariamente presentan el mismo formato. Algunas de las referencias son: *presidente Juárez*, *Juárez*, *Benito Juárez*, siendo un ejemplo de referencias de una misma EN. Estos casos de referencias de una EN es lo que se quiere aprovechar en este trabajo, utilizar el contexto de las referencias como por ejemplo Juárez y con esto apoyar la clasificación de la EN, esto llevaría a utilizar no tan sólo el contexto local de la instancia a clasificar sino considerar el contexto global. Es decir, el contexto global serán los contextos de las distintas referencias de una misma EN.

También debemos de tomar en cuenta que puede ser difícil clasificar una entidad nombrada que no ha sido vista en un ejemplo de entrenamiento o si tiene información local contextual pobre. Un ejemplo de contexto local pobre es cuando una EN se encuentra en un título o una oración corta. Considérese la siguiente oración: *Paris fue homenajeada*, donde no existe información suficiente para su clasificación, la misma oración nos sirve para ejemplificar la ambigüedad que ocurre cuando la EN se encuentra al inicio de una oración y no tiene alrededor palabras “disparadoras” que apoyen la clasificación, además de que la información que podría proporcionar la letra mayúscula inicial se pierde por ser la primera palabra en la oración, en este caso Paris puede ser un lugar, una persona o incluso una organización [1].

Pero en un conjunto de datos o en un documento distintas sentencias contienen cierta cantidad diferida de información contextual útil, donde la siguiente mención de la EN está inmersa en un contexto local útil para realizar una correcta clasificación de una EN. Esto resulta en un clasificador que predice un conjunto de etiquetas correctas e incorrectas para la misma entidad nombrada.

[1] *Este título podría referirse a la ciudad de París o a Paris Hilton, la celebridad.*

Lo que se busca en este trabajo es agrupar el contexto local de las referencias e integrar un contexto global para la clasificación y reducir el porcentaje de errores en la clasificación de una misma EN.

En este trabajo se presenta la propuesta de utilizar el contexto local de cada una de las entidades nombradas contenidas en un documento y el contexto de sus referencias para apoyar la clasificación de EN, es decir, tomar el contexto local de las coreferencias y utilizarlo junto con el contexto local de la EN para formar un contexto global. Al utilizar el contexto global de un documento es necesario contestar las siguientes preguntas: ¿cómo obtener las referencias a las ENs?, ¿cómo aprovechar la información de las distintas referencias de una misma entidad?

Para llegar a esto será necesario definir un conjunto de atributos aptos para la tarea de clasificación, definir este conjunto de atributos es relevante debido a que muchas veces el éxito o fracaso en la clasificación depende de esto. Con esto queremos mejorar el desempeño en la clasificación, estableciendo un conjunto de atributos que permitan manejar un contexto global y utilizar técnicas de aprendizaje automático para aprender a generalizar el contexto de las clases.

Una vez establecidas las tareas que se deben realizar también se debe responder ¿cuál es la mejor representación a utilizar? ¿cómo integrar los contextos locales en un contexto global?. A lo largo de este trabajo se intentará dar respuesta a las preguntas planteadas, además de resolver la propuesta inicial de mejorar el desempeño de la clasificación de entidades nombradas.

1.1.1. Objetivos

Los objetivos que se quieren alcanzar a lo largo de la investigación son los siguientes:

Objetivo general

- Diseñar un método para la clasificación de entidades nombradas basándose en información global que permita mejorar el desempeño de clasificación.

Objetivos Específicos

- Desarrollar un método de vinculación de ENs.
- Proponer un método de clasificación global basado en esquemas de votación utilizando las referencias de ENs.
- Desarrollar un método de clasificación global basado en aprendizaje automático utilizando el contexto de referencias de ENs.

1.2. Organización de la Tesis

Este documento de tesis está estructurado de la siguiente manera:

En el capítulo 2 se presentan las nociones básicas para entender el contenido del trabajo, se presentan conceptos de aprendizaje automático y el algoritmo de aprendizaje utilizado en este trabajo. También se presentan conceptos básicos en el área de Reconocimiento de Entidades Nombradas, las etapas que lo componen y sus métodos de evaluación. Posteriormente, en el capítulo 3 se hace una revisión de trabajos en el campo del REN, se revisan aquellas que han seguido una aproximación “manual” y las que hacen uso de aprendizaje automático, para posteriormente introducir los enfoques que han utilizado información local e información global.

Los capítulos 4, 5 y 6 son los más importantes de este trabajo, ya que se describe el método propuesto para resolver el problema de estudio. En el capítulo 4 se detallan las etapas que componen al método tradicional y la integración de las dos fases del método propuesto. Posteriormente, en el capítulo 5 se revisa la primera fase que involucra la vinculación de entidades nombradas, detallando los métodos de similitud propuestos y resultados obtenidos.

En el capítulo 6 se revisan los distintos enfoques propuestos utilizando información global, enfoques por voto y enfoques utilizando árboles de decisión, describiendo los resultados obtenidos para cada enfoque, además en este capítulo se hace un análisis de los resultados obtenidos por los distintos enfoques abordados.

Finalmente en el capítulo 7 se presentan las conclusiones de este trabajo y se discuten las posibles direcciones para trabajo futuro.

Capítulo 2

Conceptos Básicos

En este capítulo se introduce la teoría que fundamenta la propuesta planteada. Se presentan conceptos básicos del área de Aprendizaje Automático, como la clasificación y los árboles de decisión para familiarizar al lector con la técnica empleada en la tesis. Posteriormente en la sección 2.3 se exponen los conceptos básicos del área del Reconocimiento de Entidades Nombradas, las fases que normalmente la componen y la forma de evaluación utilizada.

2.1. Aprendizaje Automático

La tarea del Aprendizaje Automático consiste en construir programas de computadora que automáticamente mejoren con la experiencia, es decir, que aprendan. A continuación se tiene una siguiente definición más formal:

Se dice que un programa de computadora aprende de la experiencia E con respecto a un conjunto de tareas T y una medida de desempeño P, si su desempeño en tareas de T, medidas por P, mejora con la experiencia E.

Entonces un problema de aprendizaje está bien definido cuando se identifican los siguientes elementos: la clase de tareas, la medida de desempeño a mejorar y la fuente de la experiencia (ver [21] para más detalles). Distintas tareas de aprendizaje deben resolverse para lograr el objetivo de que un sistema

mejore con la experiencia, una de estas importantes tareas es la clasificación, que es una parte fundamental de este trabajo y se describirá a continuación.

2.1.1. Clasificación

Uno de los problemas fundamentales en aprendizaje es identificar miembros de clases diferentes, a lo que se llama problema de clasificación.

La clasificación es la tarea de aproximar una *función objetivo* desconocida $\Phi : I \times C \rightarrow \{T, F\}$ (donde se describe cómo las instancias deben ser clasificadas de acuerdo a un experto en el dominio) por medio de una función $\Theta : I \times C \rightarrow \{T, F\}$ llamada el clasificador, donde $C = \{c_1, ..., c_{|c|}\}$ es un conjunto de clases definido, e *I* es un conjunto de instancias del problema. Cada instancia $i_j \in I$ es representada como una lista $A = \langle a_1, a_2, ..., a_{|A|} \rangle$ de valores característicos, conocidos como *atributos*, i.e. $i_j = \langle a_{1j}, a_{2j}, ..., a_{|A|j} \rangle$. Si $\Phi : i_j \times c_i \rightarrow T$ entonces i_j es llamado un ejemplo positivo de c_i, mientras si $\Phi : i_j \times c_i \rightarrow F$ es llamado un ejemplo negativo de c_i.

El clasificador Θ es generado automáticamente mediante un proceso inductivo, llamado el *aprendiz*, el cual al observar los atributos de un conjunto de instancias preclasificadas bajo c_i o $\overline{c_i}$, adquiere los atributos que una instancia no vista debe tener para pertenecer a la clase. Por esta razón, en la construcción del clasificador se requiere la disponibilidad inicial de una colección Ω de ejemplos, tales que el valor de $\Phi(i_j, c_i)$ es conocido para cada $\langle i_j, c_i \rangle \in \Omega \times C$. A la colección de ejemplos se le llama *conjunto de entrenamiento (T_r)*. Así, al proceso anterior se le denomina *aprendizaje supervisado* debido a la dependencia de estos ejemplos de entrenamiento T_r.

Tomemos por ejemplo un problema de diagnóstico de enfermedad cardíaca donde cada instancia representa un paciente. Los atributos de cada instancia pueden incluir la edad del paciente, su sexo, su nivel de colesterol y su actividad física. En la fase de entrenamiento, además de los atributos, se recibe

la clase a la que pertenece cada ejemplo. El conjunto A corresponde a pacientes con problemas cardíacos y el conjunto B a los pacientes sin problemas cardíacos. El problema es construir con los dos conjuntos de ejemplos (ejemplos de entrenamiento) una función $f(x)$ que regrese 1 si x pertenece a A y 0 si x pertenece a B, para que cuando exista un nuevo caso de posible afección cardíaca se pueda dar un diagnóstico.

2.2. Algoritmos de Aprendizaje

Existen distintos tipos de aprendices, es decir, el proceso que permite generar automáticamente el clasificador. Diferentes algoritmos de aprendizaje se han utilizado en el área del REN, como Naive Bayes, k-Vecinos Más Cercanos, Máquinas de Vectores de Soporte, Árboles de Decisión entre otros (Ver [20] para más detalles). A continuación se explicaran de forma más específica los árboles de decisión que son la técnica empleada en este trabajo.

2.2.1. Árboles de Decisión

Los árboles de decisión son sistemas inteligentes basados en conocimiento expresado como reglas, que permiten clasificar una serie de ejemplos o salidas ordenando descendentemente los atributos que contribuyen a la clasificación en un árbol. Un árbol de decisión es un método para aproximar funciones de valores discretos, en los cuales la función aprendida se representa por un árbol de decisión. Los árboles de decisión se pueden representar como conjuntos de reglas if-then para facilitar la lectura humana. Este método de aprendizaje ha sido aplicado con éxito en una gran cantidad de tareas.

2.2.1.1. Representación de Árboles de Decisión

Los árboles de decisión clasifican instancias al ordenarlas a lo largo del árbol de la raíz a algún nodo hoja, que da la clasificación de la instancia [21]. Cada nodo del árbol especifica un atributo de la instancia y cada rama que desciende del nodo corresponde a los posibles valores del atributo. Una instancia se clasifica empezando de la raíz del árbol, probando el atributo específico en este nodo, moviéndose hacia la rama correspondiente al valor del atributo en el ejemplo dado. Este proceso se repite para el sub-árbol al siguiente nivel.

En la Figura 2.1 se muestra un ejemplo de un árbol de decisión. En el ejemplo se busca decidir si jugar o no tenis, cada instancia se clasifica recorriendo las ramas del árbol de acuerdo a los valores de los atributos de: ambiente, humedad y viento, para obtener la clasificación asociada al ejemplo, que será P o N (jugar o no jugar).

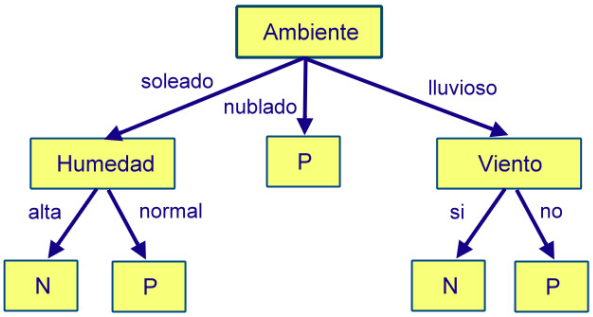

Figura 2.1: Ejemplo de árbol de decisión

Si el problema de clasificación es simple (en particular, si es linealmente separable) a menudo se prefiere utilizar árboles de decisión. Los árboles de decisión dividen el conjunto de entrenamiento en subconjuntos más pequeños, lo cual hace la generalización más difícil ya que podría no haber suficientes datos para una predicción confiable y la generalización incorrecta más fácil ya que conjuntos más pequeños tienen regularidades que no se generalizan.

2.2.1.2. Algoritmo ID3

A continuación se presenta la descripción del algoritmo ID3. Para construir el árbol, ID3 usa una aproximación descendente preguntándose al inicio, ¿qué atributo debería situarse en la raíz del árbol?, se busca encontrar el atributo más valioso en el conjunto de entrenamiento. Cada atributo es evaluado utilizando una estadística que mide que tan bien un atributo clasifica los ejemplos de entrenamiento. Una vez que se elige la raíz, se agrega una rama descendente para cada valor del atributo y los ejemplos del conjunto de entrenamiento con esos valores se ordenan en el correspondiente nodo. Se repite entonces el proceso utilizando sólo las instancias de entrenamiento asociadas a cada nodo para elegir el mejor nodo para cada nivel en el árbol. Así, para cada instancia nueva que deba ser clasificada, se evalúan sus atributos iniciando por el nodo raíz, recorriendo las ramas del árbol de acuerdo a los valores de los atributos que correspondan, se repite para cada sub-árbol encontrado hasta alcanzar un nodo hoja, donde la etiqueta de la hoja alcanzada se asigna como la clase de la instancia nueva.

En la Tabla 2.1 se describe una versión simplificada del algoritmo ID3. En el algoritmo se debe definir el atributo que mejor clasifica los ejemplos, la forma de determinar la valía de un atributo es mediante una medida estadística que es la *ganancia en la información*(GI), ésta mide qué atributo aporta mayor información separando de mejor manera las instancias del conjunto de entrenamiento conforme a las clases. En la fórmula 2.1 se presenta cómo calcular la ganancia de información $GI(S, A)$ de un atributo A relativo a un conjunto de ejemplos S, donde $Valores(A)$ es el conjunto de todos los posibles valores para el atributo A y S_v es el subconjunto de S para el cual el atributo A tiene el valor v. El primer término de la fórmula 2.1 es calculado mediante la fórmula

```
ID3(E, Ta, A)
E son los ejemplos de entrenamiento, Ta es el atributo cuyo valor es el predicho por el árbol,
A es una lista de otros atributos que pueden ser evaluados por el árbol de decisión.
- Crear un nodo raíz para el árbol
- Si todos los ejemplos son positivos. Regresar el árbol con el único nodo Raíz con etiqueta = +
- Si todos los ejemplos son negativos. Regresar el árbol con el único nodo Raíz con etiqueta = -
- Si A es vacío, regresar el nodo Raíz con etiqueta = valor más común de Ta en E
- En otro caso, Inicia
        Sea a el atributo en A que mejor clasifica E
        Etiquetar la raíz como a.
        Para cada posible valor, vi de A
            Agregar una nueva rama bajo el nodo Raíz correspondiente a la prueba a = vi
            Sea Evi el subconjunto de ejemplos E que tienen valores vi = a.
            Si Evi es vacío
                Agregar debajo de la nueva rama un nodo hoja con valor de Ta = al valor más
                 frecuente en A como etiqueta
            De lo contrario
                ID3(Evi, Ta, A-(a))
- Terminar
- Regresar raíz
```

Tabla 2.1: Algoritmo ID3

2.2.

$$GI(S,A) = Entropia(S) - \sum_{v \epsilon Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.1}$$

$$Entropia(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{2.2}$$

2.2.2. Evaluación de la Clasificación

Generalmente un clasificador se mide por su exactitud(α), la exactitud corresponde al porcentaje de decisiones correctas. Para evaluar la exactitud son de interés cuatro cantidades (Ver tabla 2.2) donde FP_i (*falsos positivos*) es el número de instancias de prueba incorrectamente clasificados como pertenecientes a la clase c_i, TN_i (*verdaderos negativos*) es el número de instancias correctamente clasificadas como no pertenecientes a la clase c_i, TP_i (*verdaderos positivos*) la cantidad de instancias clasificadas correctamente para c_i y finalmente FN_i (*falsos negativos*) es el número de instancias clasificados

incorrectamente como no pertenecientes a la clase c_i. Obteniendo estas cantidades se define la fórmula para la exactitud 2.3 .

		Respuesta Experto	
	clase c_i	SI	NO
Respuesta Clasificador	SI	TP_i	FP_i
	NO	FN_i	TN_i

Tabla 2.2: Tabla de contingencia para clase c_i

$$\alpha = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \tag{2.3}$$

Cabe mencionar que existen otras medidas para evaluar el desempeño de un clasificador, las cuales dependen de la aplicación en particular. En el contexto de este trabajo se discutirán las métricas utilizadas en el REN en la sección 2.3.1.

2.3. Reconocimiento de Entidades Nombradas

El término *Entidad Nombrada* nació en las Conferencias de Entendimiento de Mensajes (MUC) donde se buscaba promover y evaluar la investigación en el área de Extracción de Información. En las MUC se acuñó la siguiente definición:

***Entidad Nombrada**: es una palabra o secuencias de palabras que se identifican como nombre de persona, organización, lugar, fecha, tiempo, porcentaje o cantidad [27].*

La MUC funcionó como un parteaguas y a partir de entonces distintos trabajos se han hecho enfocados específicamente a la tarea de reconocimiento, debido a su utilidad en distintas aplicaciones orientadas al PLN.

El Reconocimiento de Entidades Nombradas (REN) es una tarea lingüística computacional en la cual se busca clasificar cada palabra de un documento en una determinada categoría de una lista de categorías definida, como por ejemplo pueden ser las siguientes categorías: persona, lugar, organización o fecha.

Existen distintos niveles en las tareas de extracción de información, una de ellas es la de extracción de patrones de relaciones, donde el objetivo es encontrar la relación entre pares de entidades nombradas. Por ejemplo de la frase *"Ricardo Salinas Pliego presidente de TV Azteca"*, se espera que un sistema con patrones de relaciones responda que *"Ricardo Salinas Pliego"* es una persona, *"TV Azteca"* es una organización y que Ricardo Salinas Pliego es entonces el presidente de la organización *"TV Azteca"*. Pero antes de poder establecer esta relación, es necesario clasificar correctamente los elementos *"Ricardo Salinas Pliego"* y *"TV Azteca"* como una persona y una organización respectivamente, siendo este el objetivo del REN.

La tarea del REN se resuelve generalmente como un proceso en dos pasos, los cuales se describen a continuación:

- Delimitación de Entidades Nombradas (NED): En este paso se determina la palabra o secuencia de palabras que componen a la EN. Este proceso en si es también una clasificación, donde el objetivo es clasificar cada palabra con una etiqueta de acuerdo a su pertenencia y ubicación en una EN. Para esto se sigue el esquema BIO, donde cada palabra se clasifica de la siguiente forma:
  - Una etiqueta **B** para indicar que la palabra es el inicio de la EN.
  - Una etiqueta **I** que indica su pertenencia a la EN, pero que no se encuentra al inicio de esta.
  - Una etiqueta **O** que indica que la palabra no forma parte de una EN.

Siguiendo este esquema de clasificación se tendría la frase, *"Felipe Calderón visitó la ciudad de Puebla"* etiquetada de la siguiente manera: *Felipe_B Calderón_I visitó_O la_O ciudad_O de_O Puebla_B*.

- Clasificación de Entidades Nombradas (NEC): Una vez que se han definido los límites de las ENs, se debe definir de qué tipo son, es decir, a qué clase pertenecen. Las clases definidas generalmente son: PERSONA, ORGANIZACIÓN, LUGAR Y MISCELÁNEA, donde las etiquetas para las clases son PER, ORG, LOC y MISC respectivamente; así, para el ejemplo anterior se tendría: *Felipe_B_PER Calderón_I_PER visitó_O_O la_O_O ciudad_O_O de_O_O Puebla_B_LOC*.

En la mayoría de las propuestas que quieren atacar el problema de REN, la tarea de delimitación se ve como un problema de clasificación de 3 clases (B, I, O), donde se utilizan generalmente características léxicas, sintácticas, ortográficas, afijos y características de los elementos alrededor de la palabra a clasificar. En la delimitación de las ENs son de mayor importancia atributos como la aparición de letras mayúsculas y la ubicación de las palabras en la oración. También se ha utilizado información de los vecinos (también de tipo ortográfico) para apoyar la delimitación, las propuestas van orientadas ha resolver la delimitación bajo técnicas "manuales" o automáticas. La delimitación se ha resuelto bajo técnicas manuales, basadas en reglas o patrones, o bajo técnicas de aprendizaje automático que utilizan distintos atributos para entrenar clasificadores que aprendan los patrones de la delimitación de ENs.

En el caso de la clasificación de ENs se ve generalmente como un problema de cuatro clases (ORG, LOC, PER y MISC) donde se han utilizado distintas técnicas que implican técnicas "manuales", técnicas de aprendizaje automático e incluso sistemas híbridos. Para la clasificación, los atributos a utilizar son similares a los utilizados en la primera fase pero pueden integrar algunos otros,

como palabras “disparadoras”, diccionarios, y las palabras de las ENs en si, también se utilizan atributos de los vecinos que rodean a la EN. En el paso de clasificación, atributos como la posición de las palabras en la oración e información ortográfica son menos útiles para discriminar entre clases de ENs, se consideran de forma más importante atributos como el tipo de elementos que rodean a la palabra y palabras “disparadoras” que estén muy relacionadas con alguna clase en particular.

2.3.1. Métricas de Evaluación en REN

La evaluación en Extracción de Información (EI) hace uso de las medidas de Precisión, Recuerdo y medida F, las cuales han sido adoptadas como medidas estándar en el área de REN.

Debido a que el proceso de REN se resuelve generalmente en dos pasos, la evaluación de estos se realiza de esta misma manera. Tanto para la fase de delimitación, como para la fase de clasificación se calculan las tres medidas de Precisión, Recuerdo y medida F, y los resultados se presentan como dos procesos separados.

La precisión se define como una medida de la proporción de elementos clasificados por el sistema que en realidad son correctos, se define con la fórmula 2.4.

$$p_i = \frac{tp_i}{tp_i + fp_i} \tag{2.4}$$

donde tp_i (verdaderos positivos) son los casos que el sistema clasifica correctamente para la clase c_i, mientras que fp_i (falsos positivos) son elementos clasificados dentro de una clase a la que no pertenecen.

El recuerdo mide el número de elementos correctamente identificados de una cierta clase dividido por el número total de elementos de esa clase, es decir la proporción de elementos correctos de una clase que el sistema identifica, lo

cual se expresa con la fórmula 2.5:

$$r_i = \frac{tp_i}{tp_i + fn_i} \tag{2.5}$$

donde fn_i (falsos negativos) son los elementos clasificados como no pertenecientes a una clase que si pertenecen.

En algunos casos es conveniente combinar precisión y recuerdo en una sola métrica de rendimiento global. Una forma de hacer esto es utilizando la medida F, la cual se define con la fórmula 2.6:

$$F_\alpha = \frac{(1+\alpha) * p_i * r_i}{(\alpha * p_i) + r_i} \tag{2.6}$$

donde p_i es precisión, r_i es recuerdo y α es un factor que determina la importancia que se le da a cada uno. Se elige un valor de α = 0.5 para dar un valor igual tanto a precisión y recuerdo, con este valor la fórmula se simplifica a 2.7.

$$medidaF = \frac{2 * p_i * r_i}{p_i + r_i} \tag{2.7}$$

La precisión y recuerdo pueden calcularse en cada clase c_i o mediante un promedio de todas las clases. Cuando se desea conocer la precisión y recuerdo no sólo en clases individuales sino en todo el conjunto de instancias, las medidas de evaluación se establecen de la siguiente forma; con Micro-Promedio, el cual es utilizado para dar a las categorías una importancia proporcional al número de ejemplos positivos que le corresponden, mientras que en el Macro-Promedio todas las categorías tienen la misma importancia.

Las fórmulas para Micro-Promedio son las siguientes, 2.8 y 2.9:

$$p = \frac{\sum_{i=1}^{|C|} tp_i}{\sum_{i=1}^{|C|} tp_i + fp_i} \tag{2.8}$$

$$r = \frac{\sum_{i=1}^{|C|} tp_i}{\sum_{i=1}^{|C|} tp_i + fn_i} \tag{2.9}$$

En el caso del Macro-Promedio, primero se calcula p_i y r_i para cada clase y se promedian los resultados para obtener *p* y *r* como se muestra en las ecuaciones 2.10 y 2.11.

$$p = \frac{\sum_{i=1}^{|C|} p_i}{|C|} \tag{2.10}$$

$$r = \frac{\sum_{i=1}^{|C|} r_i}{|C|} \tag{2.11}$$

Capítulo 3

Trabajo Relacionado

En este capítulo se introducen trabajos relacionados con la Clasificación de ENs, se organizan en distintos enfoques y de acuerdo a las técnicas empleadas. La organización es de la siguiente manera: primero se discuten aquellos que utilizan aproximaciones “manuales” y los que se basan en técnicas de aprendizaje automático, en la sección 3.1. En estos últimos se tiene una subdivisión, los que utilizan aprendizaje supervisado contra los que utilizan aprendizaje no supervisado. Se consideró pertinente resaltar aquellos que utilizan un contexto local para hacer la clasificación de ENs (CEN) contra los que utilizan un enfoque global, en la sección 3.1.3, esta subdivisión se realizó para distinguir las principales diferencias entre unos y otros.

3.1. Aproximaciones para CEN

Una forma de clasificar las aproximaciones que se han hecho en la tarea de clasificación de ENs es dividir las aproximaciones en “manuales” y en aquellas que se apoyan en técnicas de aprendizaje automático. Los primeros trabajos se enfocaron en construir e integrar listas de diccionarios, de palabras disparo, de reglas o patrones y obtener recursos etiquetados a mano para poder hacer el reconocimiento de EN con sistemas basados en reglas. Con el paso de los

años, el uso de técnicas de aprendizaje automático se ha vuelto popular para la resolución de problemas propios del área de PLN y por supuesto en la tarea de CEN.

En la Figura 3.1 se presenta una clasificación de la forma de atacar el problema de clasificación de ENs, las aproximaciones pueden ser manuales o automáticas, dentro de las automáticas (o de las híbridas) pueden resolverse mediante métodos supervisados o no supervisados, y finalmente se puede involucrar información local o global dando una última ramificación. A continuación se describirán algunos trabajos siguiendo esta clasificación, y posteriormente una revisión de trabajos con respecto a los enfoques que utilizan para hacer la clasificación de ENs, ya sea mediante el uso de información local o información global.

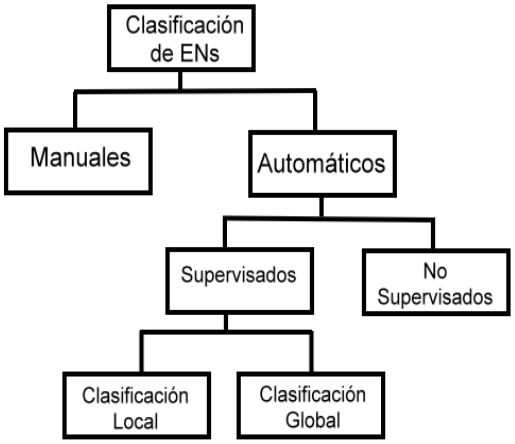

Figura 3.1: Taxonomía de trabajos de CEN

3.1.1. Aproximaciones “Manuales”

La mayoría de los sistemas participantes en el MUC-6 utilizaban aproximaciones “manuales”. Algunas características comunes de las aproximaciones se describen a continuación.

Las aproximaciones manuales son sistemas construidos a mano que recaen principalmente en la intuición de diseñadores humanos. Generalmente tienen inmerso conocimiento humano en forma de patrones o de reglas. Los patrones utilizan características gramaticales (parte de la oración), sintácticas (precedencia de la palabra) y ortográficas en combinación con diccionarios, utilizan también otros recursos como listas de palabras “disparadoras”, listas definidas de nombres de lugares, personas u organizaciones.

Por ejemplo, en uno de estos sistemas “Proteus” [12], un nombre es reconocido como de cierto tipo si está definido en un diccionario, si tiene una forma distintiva (un patrón definido) o si es un alias de un nombre de tipo conocido. Este sistema está compuesto por un gran número de reglas de producción sensitivas al contexto, bastante intuitivas. Otros ejemplos de sistemas son IsoQuest [19] y FACILE [6], sistemas similares en cuanto a que se basaban en reglas escritas manualmente y utilizaban bases de datos de ENs comunes. En estos casos se podía presentar un conflicto entre distintas reglas que predecían distintas clases para una EN, lo cual se resolvía asignando peso a las reglas. Otro ejemplo de sistema manual es el que se propone por Hobbs et al. [14], un sistema basado en expresiones regulares construidas manualmente, hace reconocimiento de frases, de patrones y resuelve incidentes. Fastus es uno de los sistemas más robustos sin embargo su mantenimiento resulta laborioso.

Estos sistemas en su mayoría alcanzan porcentajes de precisión mayores al 90 %, pero a cambio tienen un gran número de reglas que manejar (más de cien en el caso de FACILE) que cubren casos muy específicos y cuando se cambia el dominio con respecto al que fueron construidos, presentan una caída en su desempeño. Además los sistemas “manuales” presentan altos costos de mantenimiento y poca portabilidad.

3.1.2. Aproximaciones con Aprendizaje Automático

En un sistema de CEN basado en aprendizaje automático el propósito es convertir el problema de identificación en un problema de clasificación. El sistema busca los patrones y relaciones en el texto para construir un modelo utilizando algoritmos de aprendizaje automático. El sistema debe ser capaz de identificar y clasificar nombres propios en las clases particulares como personas, lugares y organizaciones basado en este modelo. Generalmente se siguen dos tipos de modelos de aprendizaje: supervisado y no supervisado.

El aprendizaje supervisado involucra utilizar un programa que pueda aprender a clasificar un conjunto de ejemplos etiquetados que están descritos por un conjunto de características. Al proceso se le dice supervisado porque las personas que etiquetan los ejemplos de entrenamiento “enseñan” al programa cómo distinguir las características importantes para diferenciar los ejemplos como pertenecientes a una determinada clase. El aprendizaje supervisado requiere preparar un conjunto de datos de entrenamiento para construir un modelo por lo cual requiere una gran cantidad de ejemplos.

En esta categoría de trabajos, Carreras, Márquez y Padró [8] presentan un sistema para la Extracción de ENs. La clasificación de ENs se aborda como un problema de clasificación de 4-clases (Organización, Lugar, Persona y Miscelánea), por lo cual se utiliza un algoritmo de aprendizaje multiclase. Se utilizan atributos léxicos y sintácticos, palabras disparo, contexto de palabras vecinas y características de diccionarios. A su vez, en [31] se estudia el problema de encontrar las ENs más importantes de un documento. Este trabajo también utiliza características de las palabras vecinas de las ENs a clasificar y aplica distintos algoritmos de aprendizaje, como Naive Bayes, Método de Minimización de Riesgos y Árboles de Decisión. Otro ejemplo de propuestas en esta dirección es el dado por Bikel et al. [4] que propuso un sistema de aprendizaje para encontrar nombres basado en Modelos Ocultos de Markov

(HMM por sus siglas en inglés) llamado Nymbel, debido a los buenos resultados obtenidos por Nymbel se propuso una versión refinada de este sistema llamado Identifinder [5], el HMM etiqueta cada palabra con una clase (persona, organización, lugar) o con una etiqueta NO-ES-NOMBRE.

A diferencia de los trabajos que se describieron anteriormente como manuales, los trabajos revisados basados en clasificación supervisada tienen rangos de precisión de 69 % a 87 %, sin embargo estos resultados se mantienen aún cuando se cambie el dominio y en algunos casos el idioma de los corpus.

En cuanto al método de aprendizaje no supervisado, su objetivo es construir representaciones de los datos, en este caso el modelo se ajusta de acuerdo a observaciones. Se distingue del aprendizaje supervisado por el hecho de que no hay conocimiento a priori.

En [22] se describe un método no supervisado que no requiere intervención humana para el etiquetamiento de los datos de entrenamiento o para la generación de diccionarios. Este sistema reconoce sólo ENs de tipo ORG, LOC y PER. El sistema utiliza un algoritmo de generación de diccionarios y se generan reglas mediante recuperación de información. Se propone un paso de desambiguación uniendo los tipos de las ENs que tengan un alias y la asignación de la clase de la EN cuyo contexto no sea ambiguo. Este sistema no supervisado no llega a ser competitivo con sistemas supervisados, al alcanzar un 70 % de valor de la medida F.

El uso de aprendizaje no supervisado no es una aproximación popular en el problema de REN debido a que no han alcanzado el nivel de desempeño de los sistemas supervisados, así que los sistemas tienden a ser sistemas híbridos que combinan módulos de aprendizaje supervisado y módulos no supervisados que generan un conjunto de reglas. Algunos ejemplos de estos sistemas se describen en [20].

Las aproximaciones basadas en aprendizaje automático tienen varias ventajas sobre los manuales. Aún cuando no tienen el conocimiento humano em-

bebido, presentan capacidad de generalización mediante los algoritmos de aprendizaje. Otra ventaja es, que no todos los idiomas son fáciles de analizar para la generación de reglas; pero si se entrena adecuadamente un clasificador con atributos útiles, aprenderá a generalizar sobre los datos de entrenamiento. Aún cuando la mayoría de los métodos requieren ser supervisados, nuevas aproximaciones se están haciendo para aprovechar técnicas de aprendizaje no supervisado y solventar la dificultad de contar con un número importante de ejemplos de entrenamiento.

3.1.3. Enfoques de Clasificación

En esta sección se presenta un análisis de trabajos enfocados a la clasificación de ENs divididos principalmente entre los que se centran en hacer el reconocimiento basados en un enfoque local (sólo utilizan atributos de la EN a clasificar) y los que presentan un enfoque global donde se utiliza información de todas las menciones de una EN, a lo que llamamos referencias a una EN.

3.1.3.1. Clasificación en un Contexto Local

En esta sección se presentan dos trabajos para el idioma español que utilizan información local como atributos para hacer la delimitación y clasificación de las ENs, la mayoría de los trabajos de CEN se encuentran orientados al idio inglés, se encuentran pocos trabajos para el español y los que se presentan aquí son los que mayor relación tienen con la propuesta de esta investigación, tanto en el corpus que utilizan como en el enfoque de mejorar la clasificación en base a una clasificación inicial. Estos métodos se basan principalmente en utilizar atributos de las ENs a clasificar (por ejemplo, etiquetas POS, información ortográfica, raíces de las palabras) y atributos de las palabras vecinas que rodean a la EN, estas palabras vecinas se encuentran en la misma oración y generalmente se establece una ventana para definir el número de elementos a

la derecha y a la izquierda que se tomarán.

En [28], Solorio y López López presentan una metodología para clasificación de EN con Máquinas de Vectores de Soporte (MVS) para mejorar la exactitud de un Sistema Extractor de ENs (SEEN). Los atributos son las clases de las palabras vecinas de la EN, etiquetas POS, información de letras mayúsculas y lemas. Para este trabajo los resultados de clasificación fueron entre 83-89 % de medida F, trabajando con las cuatro clases de Organización, Lugar, Persona y Miscelánea. En [29] se continuó con el trabajo pero manejando transcripciones orales, es decir, es independiente de signos de puntuación y capitalización. En este trabajo se concluyó que la exactitud de los sistemas se basa en su mayoría en la localización de letras mayúsculas, más que en la ubicación de signos de puntuación.

En [17] Kozareva, Bonev y Montoyo proponen explotar un gran conjunto de datos no etiquetados para mejorar la clasificación utilizando técnicas de aprendizaje semi-supervisado. En este caso este trabajo se considera importante porque fue un trabajo desarrollado y aplicado en la tarea de CEN para el idioma Español, la tarea de CEN se atacó con un método propuesto basado en *co-training* y se abarcaron las 4 clases Organización, Lugar, Persona y Miscelánea. Para la CEN se establecieron un conjunto de características locales propuestas por ellos, se trata de características léxicas y ortográficas. Para la tarea de CEN se utilizaron 792 ejemplos no etiquetados obteniendo un valor de medida F de 67.22 %.

3.1.3.2. Clasificación en un Contexto Global.

En los siguientes trabajos se presentan aquellos que siguen un enfoque global, es decir que utilizan información no sólo de la EN que se quiere clasificar sino de otras menciones de la misma EN, a estas menciones de la misma EN se le llaman referencias y se pueden obtener distintos atributos de ellas; por ejemplo, información ortográfica, ubicación en la oración, raíz de la pala-

bra. Estas distintas menciones de una misma EN y sus atributos es lo que se denomina información global. Algunos trabajos que utilizan información global se describen a continuación.

En el trabajo de Chieu y Ng [9] se describe el sistema MENERGI (Maximum Entropy Named Entity Recognizer using Global Information), en este trabajo se utiliza información de todo el documento para clasificar cada palabra con sólo un clasificador. MENERGI utiliza características locales y globales, las primeras se toman de las palabras vecinas y la palabra en si. Las características globales se extraen de otras ocurrencias del mismo token en todo el documento. El clasificador está basado en Máxima Entropía y se centra en las clases de Organización, Lugar, Persona y Miscelánea. Chieu y Ng realizaron distintas pruebas agregando características globales alcanzando resultados de 87.24 % para MUC-7 y 93.27 % para MUC-6 para el valor de la medida F.

En el trabajo de Ji y Grishman [15] se presenta un método de resolución de coreferencias utilizando información global, para el idioma Chino. Se realiza análisis de coreferencia y se utiliza evidencia de otras frases para desambiguar nombres, se procesan documentos similares conteniendo evidencia de instancias adicionales. Se utiliza un clasificador MVS entrenado con un conjunto de características de referencias para clasificar entidades que en el primer paso no pasaron un umbral de clasificación. Un aspecto a considerar de este trabajo y que representa una desventaja es el número de pasos que realizan para la clasificación y el conjunto de reglas que se deben de establecer. Aún con esto los resultados son buenos obteniendo incluso un valor de precisión de 92.2 %, recuerdo de 89.6 % y de 90.9 % para la medida F, en el contexto de este trabajo se busca clasificar ENs de tipo Organización, Persona y Entidad Geopolítica.

En [30], Wong y Ng plantean un método para explotar texto no etiquetado. Se establece una propiedad empírica, que es “muchas ENs ocurren sólo en una clase”, se asume que las distintas apariciones en un documento de una misma EN pertenecerán a una sola clase, por ejemplo, si la EN París aparece

tres veces, en las tres ocasiones se le asignará la etiqueta LOC. Las características globales toman características de sentencias conteniendo a la EN, las características son ortográficas, léxicas y otras, como palabras clave o sufijos. Se obtiene una *etiqueta mayoritaria*, entrenando un clasificador (basado en máxima entropía) con esta etiqueta. Los resultados reportados son de 87 % para medida F. Los resultados en este trabajo son buenos, aunque hay que considerar que los autores trabajan con textos en inglés, y ellos mismos mencionan que con el idioma inglés se cumple la propiedad mencionada en un 90 %, a diferencia del idioma español donde la propiedad se cumple sólo en un 76 %.

En las secciones anteriores se presentaron una serie de trabajos relacionados con la clasificación de ENs, en la Tabla 3.1 se presenta una serie de características de manera resumida de cada uno de estos.

Sistema	Clasif.	Idioma	Corpus	Clases	Medida F
Proteus [12]	Manual	Inglés	MUC-6	ENAMEX-TIMEX-NUMEX	91.0 %
IsoQuest [19]	Manual	Inglés	MUC-7	ENAMEX-TIMEX-NUMEX	91.6 %
Facile [6]	Manual	Inglés	MUC-7	ENAMEX-TIMEX-NUMEX	82.25 %
Carreras [8]	Supervisado	Inglés	CoNLL-2002	ORG-PER-LOC-MISC	85.0 %
Zhang [31]	Supervisado	Chino	Noticias	PER	77.82 %
Bikel [4]	Supervisado	Inglés	MUC-6	PER	94.9 %
Identifinder [4]	Supervisado	Inglés	MUC-7	PER-ORG-LOC	93.6 %
Nadeau [22]	No superv.	Inglés	MUC-7	ORG-PER-LOC	70.0 %
Kozareva [27]	C.Local	Español	CoNLL-2002	ORG-PER-LOC-MISC	67.22 %
Solorio [27]	C.Local	Español	CoNLL-2002	ORG-PER-LOC-MISC	89 %
MENERGI [9]	C.Global	Inglés	MUC-6	ORG-PER-LOC-MISC	93.27 %
MENERGI [9]	C.Global	Inglés	MUC-7	ORG-PER-LOC-MISC	87.24 %
Grishman [15]	C.Global	Chino	MUC-7	PER-ORG-GPE	90.9 %
Wong [30]	C.Global	Inglés	ConLL-2003	ORG-PER-LOC-MISC	87 %

Tabla 3.1: Resumen de trabajos de CEN

En esta sección se han presentado algunos de los trabajos con mayor relación a nuestra propuesta, donde se utilizará información global y por lo cual entra en esta última clasificación. De la revisión podemos hacer el siguiente

análisis:

- El enfoque local presenta una mayor simplicidad al integrar menos información para realizar la clasificación, sin embargo, las propuestas que utilizan información local de la EN a clasificar no resuelven el problema de ambigüedad que ocurre en un contexto pobre y que podría solventarse utilizando información global.

- Las propuestas de información global descritas agregan las referencias de una EN como atributos para la clasificación, sin integrar completamente el contexto de estas referencias, es decir, no se busca integrar toda la información que se pueda obtener.

- El enfoque global presenta la ventaja de poder utilizar mayor cantidad de información disponible para apoyar la clasificación, aunque esto presentará un mayor trabajo computacional.

- Cabe mencionar que no se encuentran reportados trabajos para el idioma español que utilicen algún tipo de información global para la clasificación, sólo se han reportado trabajos que trabajan con un contexto local, es por esto que se creé importante utilizar información global para analizar el impacto que pudiera tener en el idioma español.

Nuestro enfoque difiere de los anteriores básicamente en que se propone hacer uso de los contextos locales de las distintas referencias (como una forma de información global), es decir no sólo tomar en cuenta la información que brinda la coreferencia por si misma sino considerar la información de la coreferencia y sus elementos vecinos para poder integrarla como una forma de información global para la entidad nombrada a clasificar. También en este trabajo se busca revisar distintos métodos de vinculación de ENs, para definir el adecuado para la clasificación y utilizar los elementos de esta vinculación como atributos. Además se propondrán distintas formas de integración en cuanto

a los contextos locales, utilizando un enfoque de voto (local y ponderado) así como enfoques utilizando árboles de decisión para integrar las referencias de una EN como atributos. Cabe mencionar que este trabajo está enfocado al i-dioma español, donde no se ha utilizado anteriormente información global para la CEN; las clases con las que se trabajaran son Organización, Persona, Lugar y Miscelánea, que son con las que se trabajan generalmente en la clasificación de ENs, además de que los corpus disponibles se encuentran etiquetados bajo las clases mencionadas.

Capítulo 4

Arquitectura del Método Propuesto

En los capítulos anteriores se ha planteado la problemática y los distintos enfoques que se han seguido en otras propuestas para solventar las dificultades que presenta la clasificación de ENs, con base en esto se plantea integrar un contexto global para apoyar la clasificación de ENs. En este capítulo se describe la arquitectura general del método propuesto. En la sección 4.1 se detallan los elementos que componen al método propuesto y la función que desempeña cada uno de los bloques. En los capítulos posteriores, 5 y 6, se explican las dos fases medulares de la propuesta.

4.1. Arquitectura General

Como se describió en la sección 2.3, el proceso de REN se resuelve generalmente como un proceso en dos fases, la delimitación de ENs y su posterior clasificación. El trabajo propuesto consiste en un método para refinar la clasificación inicial de ENs que pertenezcan a cadenas de referencias de una misma EN. Con este paso de refinamiento se busca corregir aquellas clasificaciones de ENs que sean incorrectas y que al mismo tiempo pertenezcan a una cadena de referencias, para así, poder aprovechar la información de esta cadena. Tomemos el siguiente ejemplo que integra la clasificación inicial las ENs que

se refieren a *Benito Juárez*:

En 1861, después del triunfo liberal de la Guerra de Reforma, el gobierno de ***Juárez-ORG*** *se instaló en la capital de la República. La familia Juárez-Maza, vivió en aquel lugar los momentos de mayor intimidad doméstica, disfrutando de la paz que* ***Juárez-PER*** *había logrado para toda la nación. El 18 de julio de 1872,* ***Benito Juárez-PER*** *fallecía en la que había sido su habitación conyugal, en la casa de su familia que se convertiría, años después, en el recinto de homenaje a su memoria.*

Se puede observar que en la clasificación inicial hay un elemento incorrecto, en la frase "*el gobierno de* ***Juárez***" la EN es clasificada de tipo organización cuando le corresponde la clase persona, sin embargo, se puede observar que los otros dos elementos son clasificados correctamente. La propuesta de este trabajo se basa en aprovechar precisamente la información de los elementos correctos para corregir la clasificación del primer elemento.

En este trabajo se presenta un proceso en cuatro pasos que involucran las dos fases clásicas del REN y dos fases parte de la propuesta: la delimitación de ENs, la clasificación inicial, la vinculación de ENs y el refinamiento de la clasificación inicial.

En la Figura 4.1 se presenta un esquema general de los pasos que se siguen para obtener la clasificación final de las ENs de un documento.

A continuación se describen cada uno de los bloques que componen a la Figura 4.1, los primeros dos pasos componen a la clasificación inicial cuyo propósito es identificar las ENs y obtener una primera clasificación de estas. Los dos pasos finales son los que involucran al método propuesto, donde se busca vincular todas las ENs que tengan referencias a una misma EN y utilizar esta vinculación para modificar la clasificación final y en su caso corregir aquellos errores que existieran en la primera clasificación.

- **Delimitación de ENs**. El primer paso de delimitación tiene como propósito identificar las palabras que forman parte de una EN y delimitar cada

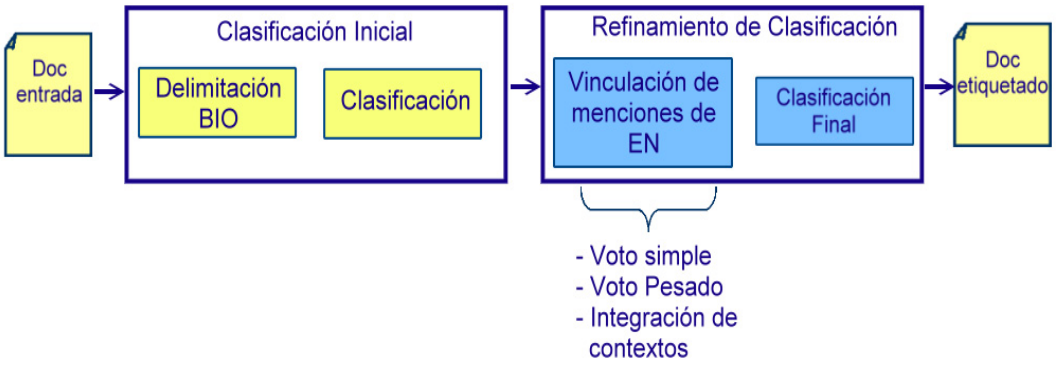

Figura 4.1: Método propuesto

una de las ENs pertenecientes a un documento. En este primer paso se recibe un documento de entrada, el cual es procesado por una herramienta de delimitación que genera un nuevo documento etiquetado bajo el esquema BIO (descrito en la sección 2.3). Este documento etiquetado de salida contiene la delimitación de las entidades nombradas encontradas. Tomemos las siguientes oraciones de un documento, que contiene la EN **Victoria** en distintas partes del documento, la salida de la delimitación sería de la siguiente forma:

Un-O juez-O del-O Tribunal-B Supremo-I del-O estado-O de-O ***Victoria-B*** *.-O El-O Abogado-B General-I de-O* ***Victoria-B*** *indicó-O que-O no-O se-O puede-O controlar-O la-O información-O .-O Juzgado-O en-O la-O ciudad-O de-O Melbourne-B capital-O de-O* ***Victoria-B*** *.-O*

- **Clasificación inicial**. Como se mencionó anteriormente se busca mejorar la clasificación de ENs, por lo cual, es necesario obtener una primera clasificación inicial; que es el propósito de este segundo paso. Con esta primera clasificación se busca tener un primer indicio de la clase correcta de las ENs de un documento y hacer uso de la primera clasificación

1 Un uno DI O O	*16 no no RN O O*
2 juez juez NC O O	*17 se haber VA O O*
3 del SP O O	*18 puede poder VM O O*
4 Tribunal tribunal NP B NP00O00	*19 controlar controlar VM O O*
4 Supremo supremo NP I NP00O00	*20 la el DA O O*
5 del de SP O O	*21 información información NC O O*
6 estado estado NC O O	*22 . . Fp O*
7 de de SP O O	*23 Juzgado juzgar VM O O*
8 Victoria victoria NP B NP00G00	*24 en en SP O O*
9 . . Fp O O	*25 la el DA O O*
10 El el DA O O	*26 ciudad ciudad NC O O*
11 Abogado abogado NP B NP00S00	*27 de de SP O O*
11 General general NP I NP00S00	*28 Melbourne melbourne NP B NP00G00*
12 de de SP O O	*29 capital capital AQ O O*
13 Victoria victoria NP B NP00O00	*30 de de SP O O*
14 indicó indicar VM O O	***31 Victoria victoria NP B NP00G00***
15 que que CS O O	*32 . . Fp O*

Tabla 4.1: Salida de clasificación inicial

como un atributo de información a ser aprovechado. En el segundo paso se obtiene una primera clasificación, que llamamos clasificación inicial, esta es la clasificación que se toma como base y que se busca refinar en una segunda clasificación. Además de la clasificación inicial se obtiene un conjunto de atributos, como la raíz de la palabra, su POS (Parte de la oración) y numeración. Por ejemplo, para el párrafo mencionado anteriormente se tiene la salida que se muestra en la Tabla 4.1.

El número al inicio de cada línea indica la numeración de los elementos en el documento (una EN se toma como un único elemento, por lo

que se repite el número para cada palabra que la compone), el segundo elemento es la palabra, seguida por la raíz de la palabra, su etiqueta POS (Part-of-speech o parte de la oración), la etiqueta BIO (obtenida en el primer paso) y finalmente la etiqueta de clasificación. Para la etiqueta de clasificación se tiene como salida lo siguiente: *NP00SP0* que indica la etiqueta para Persona, *NP00S00* etiqueta para Miscelánea, *NP00G00* etiqueta para Lugar, *NP00O00* etiqueta para Organización.

Además de los atributos mencionados se obtiene un valor de confianza (v_{conf}) cuando la clasificación inicial es asignada por la herramienta FreeLing [1] (para más detalles ver sección 6.1.3.1), este valor indica la confianza con que se asigna una determinada clase a cada EN. Los valores de confianza pueden ser positivos o negativos, para la clase Miscelánea se obtienen valores negativos que indican que la EN no pudo ser etiquetada como alguna clase del tipo Persona, Lugar y Organización, para estas tres clases los valores de confianza son positivos.

- **Vinculación de entidades**. La fase de vinculación de ENs tiene como objetivo enlazar todas las ENs que hacen referencia a una misma EN, esta agrupación permitirá obtener información no sólo de la EN a clasificar sino de todas las ENs que hacen referencia a ella y que será aprovechada para el último paso de clasificación. Después de la clasificación inicial, se vinculan las ENs que son referencia de una misma EN para obtener como salida un conjunto de EN contenidas en cadenas. En la Tabla 4.2 se ilustra el tercer paso de vinculación. Este proceso se describe en el capítulo 5 con más detalle.

- **Clasificación Final**. En este último paso se busca corregir los errores que se tienen en la clasificación inicial, haciendo uso de la información

[1] *FreeLing Home Page. http://garraf.epsevg.upc.es/freeling/index.php*

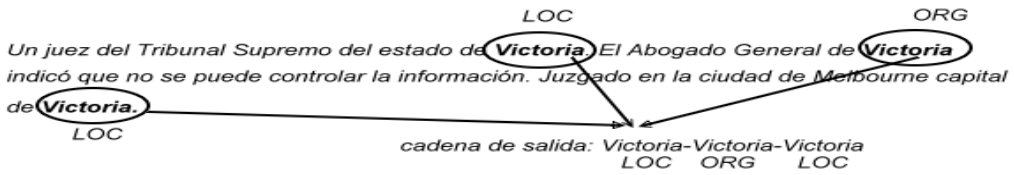

Tabla 4.2: Salida de vinculación

proporcionada por el paso de vinculación de entidades, es decir el propósito es utilizar la información de cada EN perteneciente a una cadena para corregir posibles clasificaciones erróneas. Esta última fase recibe como entrada la clasificación inicial con los atributos mencionados y las cadenas de EN, vinculadas en la fase anterior. En este trabajo se abordan varios enfoques para asignar la clasificación final, como enfoques basados en el voto de las clases asignadas y enfoques basados en árboles de decisión (descritos en el capítulo 6) y se utilizan atributos de las cadenas obtenida para integrarlos como características o factores de decisión para asignar la clase final. Para el ejemplo utilizado a lo largo de esta sección se tendría la salida mostrada en la Tabla 4.3. Es decir con los enfoques propuestos buscamos corregir los errores de clasificación de las ENs que se encuentren encadenadas.

Un juez del Tribunal Supremo del estado de ***Victoria-LOC****. El Abogado General de* ***Victoria-LOC*** *indicó que no se puede controlar la información. Juzgado en la ciudad de Melbourne capital de* ***Victoria-LOC.***

Tabla 4.3: Salida de clasificación final

Capítulo 5

Vinculación de Entidades Nombradas

En este capítulo se expone la primera fase agregada al método tradicional de CEN, la vinculación de entidades, donde se busca enlazar las distintas referencias de una EN. En la sección 5.1 se describe el primer paso para encontrar las referencias a ENs y las medidas de similitud utilizadas para establecer un vínculo entre las ENs, posteriormente se explica la forma de evaluación y resultados obtenidos (sección 5.2). En el siguiente capítulo se expone de forma detallada la fase 2 que involucra la fase medular de la propuesta de este trabajo de tesis.

5.1. Método de Vinculación

El método de vinculación de entidades nombradas tiene como objetivo encontrar las distintas menciones que se hagan de una misma entidad nombrada, para formar un conjunto de entidades referenciadas y así aprovechar la información de cada una de las apariciones de la entidad en un documento determinado, al conjunto de entidades referenciadas se le llama “cadena de

referencias".

El método de vinculación busca agrupar las ENs que posean cierta similitud entre ellas, para poder ser utilizadas en el siguiente paso de clasificación. En el método de vinculación para cada una de las ENs se encuentran sus referencias, es decir se encuentran todas las ENs que presenten similitud entre ellas, esta similitud permitirá que las ENs sean agrupadas en un conjunto de cadenas y poder utilizar la información, no sólo de las ENs como un elemento sino de las cadenas como un conjunto de información.

Antes de realizar la vinculación de ENs se realiza un preprocesamiento a las ENs clasificadas: la expansión de siglas y acrónimos, el objetivo de este paso de preprocesamiento es encadenar un mayor número de ENs y por lo tanto encontrar un mayor número de referencias. El proceso de expansión se basa en lo siguiente; básicamente si se encuentra un elemento que esté compuesto sólo de letras mayúsculas se busca este elemento en una lista de siglas definida y se obtiene el significado de dicha sigla. La lista de acrónimos fue proporcionada por miembros del Laboratorio de Tecnologías del Lenguaje (LabTL) del INAOE y obtenida de manera automática de un conjunto de noticias de idioma español de la agencia Efe [10].

Para hacer la vinculación se definen dos reglas, descritas en la Tabla 5.2, estas reglas permiten poner restricciones para evitar posibles errores en la vinculación. La primera regla se establece para restringir la vinculación de ENs que estén incluidas en otras pero que no deban vincularse, por ejemplo, si se tiene la EN *Ayuntamiento de Puebla* y la EN *Puebla* en un documento, éstas entidades no se refieren a la misma EN, y por lo tanto no deben vincularse, aún cuando la EN *Puebla* se encuentre incluida totalmente en la EN *Ayuntamiento de Puebla*. Así, esta primera regla se basa en descartar elementos de este tipo y que puedan introducir elementos erróneos a la cadena de referencias.

La segunda regla consiste en sólo aceptar elementos que sean de menor longitud (con respecto al número de palabras que conforman la EN) que el

primer elemento que se agregó a la cadena de referencias, esta regla se establece debido a que la mayoría de las ocasiones la primera mención de una EN se se hace de manera formal, es decir, si se habla de una persona lo más probable es que se escriba su nombre completo y posteriormente se utilice su apellido o nombre de pila para referirse a esta persona.

El método inicia cuando se tiene como entrada el documento con las EN delimitadas y clasificadas, básicamente el primer paso consiste en tomar cada una de las ENs de un documento y medir la similitud que existe entre ésta y las demás ENs, si existe similitud y se cumple entre las ENs la regla 1 de la Tabla 5.2 entonces se debe agregar el índice correspondiente en la EN, en este primer paso el objetivo es encontrar las ENs que presenten similitud entre ellas e irlas referenciando mediante un índice, (en la sección 5.1.1 se describen con más detalles los métodos para medir la similitud entre ENs).

Una vez que se encuentran las referencias, se busca agrupar en una cadena aquellas que presentaron cierta similitud en el paso anterior, para generar un conjunto de cadenas, que en primera instancia está vacío. Se guarda entonces temporalmente la primera EN en una cadena y se analizan cada una de las EN que se referenciaron en el primer paso, las ENs se agrupan siguiendo la regla 2 (ver Tabla 5.2), si se cumplen las condiciones se procede a agregar la EN a la cadena de referencia y se marca como perteneciente a dicha cadena para no volver a agregarla a otra cadena, el método de vinculación se encuentra descrito en la Tabla 5.1.

5.1.1. Medidas de Similitud

En esta sección se describen las medidas de similitud utilizadas, las cuales permitirán medir que tan parecidas son las ENs entre ellas de forma léxica y así poder establecer una vinculación entre un conjunto de ENs.

Se analizaron cuatro métodos para determinar la similitud entre las ENs del

```
Vinculacion(ENs)
Sea ENs el conjunto de entidades nombradas por documento
Sea numENs el número de entidades nombradas por documento
- Definir valormin de similitud
- Para toda ENi tal que 1 < i < numENs hacer
     Si ENi es acrónimo entonces
          ENi ← expansion de ENi
- Crear cadena_temp
- Para toda ENj tal que 1 < j < numENs y i!=j hacer
     Calcular similitud entre ENi y ENj
     Si similitud > valormin entonces
          indice = j
          Si indice no se ha agregado entonces
             Agregar indice a ENi[indices]
             Agregar ENj a cadena_temp
             Marcar ENj agregada
          De lo contrario descartar ENj
     De lo contrario descartar ENj
- Si |cadena_temp| > 0 entonces
     Agregar cadena_temp a lista_cadenas
- De lo contrario descartar cadena_temp
- Terminar
```

Tabla 5.1: Algoritmo de vinculación de ENs

Regla 1. *Sea $EN_{inicial}$ la primera EN de la cadena y de tipo $< palabra1 >$ de $< palabra2 >$ y sea EN_i una EN subsecuente sólo compuesta por $< palabra2 >$, entonces, no vincular EN_i con $EN_{inicial}$.* Por ejemplo, *<Museo> de <España>* no se vincularía con la EN *<España>*
Regla 2. *Agregar EN_i a la cadena sólo si $\|EN_{inicial}\| > \|EN_i\|$.*

Tabla 5.2: Reglas de vinculación

documento y asignar la referencia entre ENs. De estos cuatro métodos dos se basaron en medidas de similitud, que son una clase de medidas que permiten medir que tan parecidas son un par de cadenas (en este caso las palabras que componen a la EN) y que obtienen como resultado un valor de similitud o disimilitud entre el par de ENs. Los otros dos métodos se encuentran basados en la contención de una EN en otra.

- Similitud exacta. Con esta aproximación se realiza un emparejamiento exacto de las entidades nombradas, es decir todos los elementos que componen a una EN deben ser iguales a los que componen a la EN con la que se compara. Se define de la siguiente manera, donde EN_i y EN_j son las ENs entre las que se va a medir la similitud:

  Sea $EN_i = P_1, ..., P_n$ y $EN_j = Q_1, ..., Q_m$ donde n = número de elementos de EN_i y m = número de elementos de EN_j.

  Si m=n entonces

$$Sim_{ex}(EN_i, EN_j) = \begin{cases} 1 & \text{si} \quad \forall k, P_k = Q_k \\ 0 & \text{otro caso} \end{cases}$$

- Similitud Dice. La medida de coeficiente Dice es una medida de similitud basada en términos que se define como el doble del número de términos

comunes de las entidades comparadas dividido entre el total de número de términos en las dos entidades evaluadas. El resultado del coeficiente de 1 indica elementos idénticos. Se define la medida de similitud con la fórmula 5.1.

Sea $EN_i = \{P_1, ..., P_n\}$ *y* $EN_j = \{Q_1, ..., Q_m\}$ *donde n = número de elementos de* EN_i *y m = número de elementos de* EN_j.

$$Sim_{Dice}(EN_i, EN_j) = \frac{2 \times |EN_i \cap EN_j|}{|EN_i| + |EN_j|} \tag{5.1}$$

Donde EN_i y EN_j son las entidades nombradas a comparar. En nuestro caso establecimos un umbral para el valor de similitud que se obtiene del coeficiente Dice, si el valor es mayor a 0.5 se asume que existe una similitud entre los elementos y se agrega el índice de la EN (EN_j) como referencia de la EN (EN_i)con la que se está comparando.

- Similitud de Superposición. Esta es una medida donde si un conjunto A es un subconjunto de B o viceversa el coeficiente de similitud es 1. Se define con la fórmula 5.2, en este caso se compara si los elementos de una EN son subconjunto de la EN con la que se compara.

  Sea $EN_i = \{P_1, ..., P_n\}$ *y* $EN_j = \{Q_1, ..., Q_m\}$ *donde n = número de elementos de* EN_i *y m = número de elementos de* EN_j.

$$Sim_{Sup}(EN_i, EN_j) = \frac{|EN_i \cap EN_j|}{min(|EN_i|, |EN_j|)} \tag{5.2}$$

- Contención exacta de una cadena en otra. En esta medida se busca que uno o más elementos de EN_i estén contenidos en EN_j respetando el orden de los elementos, es decir que los elementos de EN_i sean un subconjunto ordenado de EN_j.

  Sea $EN_i = P_1, ..., P_n$ *y* $EN_j = Q_1, ..., Q_m$ *donde n = número de elementos de* EN_i *y m = número de elementos de* EN_j

$$Sim_{ConEx}(EN_i, EN_j) = \begin{cases} 1 & \text{si} \quad P_k = Q_{k+s},\ k \leftarrow 0\, a\, n, si\, k+s < m \\ 0 & \text{otro caso} \end{cases}$$

5.2. Evaluación de Vinculación de EN

Medir el desempeño de un sistema es un aspecto de suma importancia para un algoritmo de coreferencias. En el contexto de este proyecto no se resuelven todos los tipos de coreferencias que puedan existir, sin embargo si se busca enlazar las referencias a una misma EN con otras ENs que presenten similitud (definidas en las medidas de similitud), por esta razón se consideró importante definir una forma de evaluación para medir las cadenas obtenidas siendo el enfoque de evaluación de coreferencias el más adecuado y con el que se presentaron mayores semejanzas con el método de vinculación propuesto.

5.2.1. La Métrica de Evaluación B-CUBED

En [3] se describe y analiza un algoritmo de medición de coreferencia, utilizado para evaluar los sistemas de coreferencia de la sexta Conferencia de Entendimiento de Mensajes (MUC-6), la métrica B-CUBED, que es la que se utiliza en este trabajo para medir el desempeño de las medidas de similitud propuestas.

Para una entidad, i, se define la precisión y recuerdo con respecto a la entidad como se muestra en las fórmulas 5.3 Y 5.4, donde cadena $salida_{correctosi}$ se refiere al número de elementos correctos en la cadena de salida que contiene a la $entidad_i$, y $salida_i$ corresponde al número de elementos en la cadena de salida conteniendo a la $entidad_i$, para la fórmula de recuerdo, $original_i$ se refiere al número de elementos en la cadena verdadera conteniendo a la

$entidad_i$

$$Precision_i = \frac{\#\ salida_{correctosi}}{\#\ salida_i} \tag{5.3}$$

$$Recuerdo_i = \frac{\#\ salida_{correctosi}}{\#\ original_i} \tag{5.4}$$

Así la precisión y recuerdo final se calculan con las siguientes dos fórmulas:

$$Precisión\ Final = \sum_{i=1}^{N} w_i * Precisión_i \tag{5.5}$$

$$Recuerdo\ Final = \sum_{i=1}^{N} w_i * Recuerdo_i \tag{5.6}$$

donde N es el número de entidades en el documento, y w_i es el peso asignado a la entidad i en el documento, en nuestro caso se darán pesos iguales para cada entidad.

Para entender de manera clara la forma de evaluación en que se basa B-Cubed, se describe a continuación un ejemplo. Supongamos que se tienen tres cadenas de ENs, que son cadenas de las ENs *Fox, Grupo Fox y Vicente Fox*, asumiendo que Fox se refiere a la cadena de televisión, Grupo Fox a una asociación civil y Vicente Fox al ex-presidente de México, y que estas ENs se mencionan a lo largo de un documento. Por ejemplo, que la EN *Fox* se menciona 4 veces en el documento, la EN *Grupo Fox* 2 veces y la EN *Vicente Fox* 5 veces. En la Figura 5.1 se muestra la forma en que deben ser encadenadas las ENs, las filas etiquetadas como *cadena 1, cadena 2 y cadena 3* pertenecen al conjunto de cadenas de salida de las ENs Fox, Grupo Fox y Vicente Fox, la flecha abajo de cada EN representa el vínculo que se establece entre cada EN, es decir, que cada EN es referencia de las demás menciones de esta misma EN.

En la figura 5.2 se muestra un ejemplo, suponiendo que el método de vinculación uniera las cadenas que contienen las ENs *Grupo Fox y Vicente*

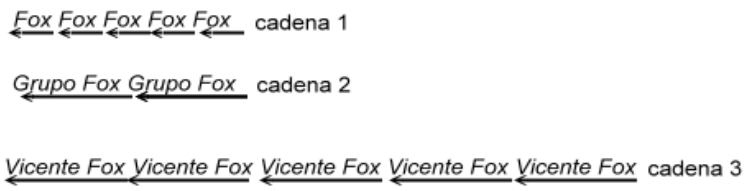

Figura 5.1: Cadenas verdaderas

Fox, se tendrían 7 elementos encadenados en una cadena, que serían las menciones de *Grupo Fox Grupo Fox* con *Vicente Fox Vicente Fox Vicente Fox Vicente Fox Vicente Fox Vicente Fox*. Para este ejemplo el método de vinculación lo vemos como una caja negra, lo que nos importa es la forma en que se evalúan las cadenas de salida cuando se tienen elementos erróneos en las cadenas.

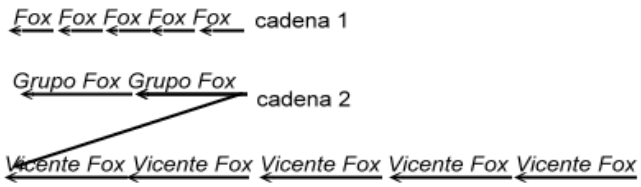

Figura 5.2: Ejemplo cadenas de salida 1

En las figura 5.3 se tiene otro ejemplo, donde se unen las ENs *Fox y Vicente Fox*, en este caso la cadena de salida tendría 10 elementos, es decir, la cadena sería *Fox Fox Fox Fox Vicente Fox Vicente Fox Vicente Fox Vicente Fox Vicente Fox*. Podemos ver que el número de elementos erróneos encadenados es mayor que en el ejemplo anterior donde el tamaño de la cadena de salida es menor, esto debería de verse reflejado al momento de realizar la evaluación, es decir, es importante penalizar la introducción de elementos que no pertenecen a la cadena original y favorecer aquellas que no presenten estos elementos erróneos.

En el trabajo de Bagga y Baldwin [3] se afirma que cuando se utiliza la coreferencia para una tarea de extracción de información donde la información

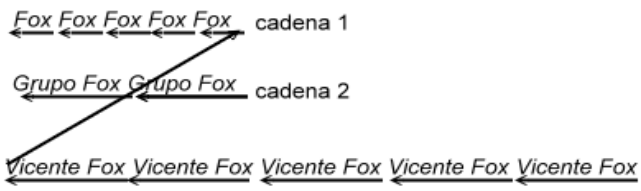

Figura 5.3: Ejemplo cadenas de salida 2

acerca de cada entidad en una clase equivalente es importante, el esquema de pesado asigna pesos iguales para cada entidad i, por esta razón el valor de w_i es de $1/12$. Por otro lado se puede ver que los valores de precisión para las respuestas en las Figuras 5.2 y 5.3 son de 76 % y 58 % respectivamente, mientras que los valores de recuerdo son de 100 %.

En la tabla 5.3 se muestran los valores de precisión y recuerdo para los ejemplos anteriores. Las filas 1 y 2 son los valores para el ejemplo 1 (dónde se encadenan *Grupo Fox y Vicente Fox*), y las filas 3 y 4 corresponden a los valores para el ejemplo 2 (dónde se encadenan Fox y Vicente Fox). En el ejemplo 1 se observa que el número de elementos encadenados es menor que en el ejemplo 2, y por lo tanto se espera que haya una penalización distinta para distinguir este tipo de casos erróneos, eso se aprecia en la Tabla 5.3 con los resultados de precisión para los dos ejemplos, siendo el segundo ejemplo (aquel que enlaza más elementos erróneos) el que recibe menores resultados. Es decir, este método penalizará de forma más dura el encadenar un mayor número de elementos erróneos y por lo tanto las medidas de similitud que realicen vinculaciones de este tipo se verán menos favorecidas en los resultados de precisión.

Salida	Medida B-Cubed (peso igual para cada entidad)
Ejemplo 1	$P: \frac{1}{12} * \left[\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{2}{7}+\frac{2}{7}+\frac{5}{7}+\frac{5}{7}+\frac{5}{7}+\frac{5}{7}+\frac{5}{7}\right] = \frac{16}{21}\,(76\,\%)$ (5.7)
	$R: \frac{1}{12} * \left[\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{2}{2}+\frac{2}{2}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}\right] = (100\,\%)$ (5.8)
Ejemplo 2	$P: \frac{1}{12} * \left[\frac{5}{10}+\frac{5}{10}+\frac{5}{10}+\frac{5}{10}+\frac{5}{10}+\frac{2}{2}+\frac{2}{2}+\frac{5}{10}+\frac{5}{10}+\frac{5}{10}+\frac{5}{10}+\frac{5}{10}\right] = \frac{7}{12}(58\,\%)$ (5.9)
	$R: \frac{1}{12} * \left[\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{2}{2}+\frac{2}{2}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}+\frac{5}{5}\right] = (100\,\%)$ (5.10)

Tabla 5.3: Valores de la métrica B-Cubed de ejemplos 1 y 2

5.2.2. Resultados de Vinculación de Entidades

Una vez que se estableció la forma de evaluación de las cadenas obtenidas se realizaron pruebas con 250 documentos del conjunto de datos de entrenamiento del corpus CoNLL (para más detalles del corpus ver capítulo 6).

Este conjunto de documentos sólo contienen las etiquetas propias de la delimitación y clasificación de las EN, no contienen etiquetas que indiquen las referencias entre las ENs y por lo tanto se dificulta la comparación y evaluación de las cadenas obtenidas. Por esta razón se etiquetaron 250 documentos manualmente con las referencias entre entidades nombradas y es contra estos

250 documentos etiquetados manualmente que se comparan las cadenas de ENs obtenidas.

En el etiquetado manual a cada EN se le asigna el índice de la cadena a la que pertenece. En la Figura 5.4 se muestra la forma en que se asigna cada índice, en este ejemplo se observa que las entidades “Abogado General del Estado” y “Abogado General´´ pertenecen a la misma cadena, que es la número 1 en una lista de cadenas, a su vez las entidades de “CrimeNet” pertenecen a la cadena número 7. De esta forma se van vinculando mediante el índice las ENs que presentan similitud entre ellas.

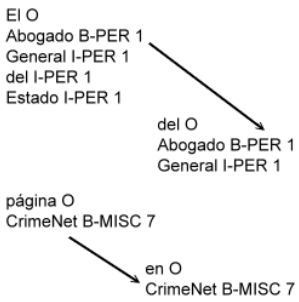

Figura 5.4: Ejemplo de vinculación

Una vez que se tienen estas cadenas etiquetadas se realizó la evaluación de cada uno de los métodos utilizados para comprobar cuál de ellos es el mejor para ser utilizado en la segunda fase. Se evaluaron los métodos utilizando la métrica descrita en 5.2.1. Los resultados se muestran en la Tabla 5.4.

Médida	**P(%)**	**R(%)**	**Correctas**	**Incorrectas**
Sim_{Ex}	**92.17**	81.20	767	65
Sim_{Dice}	90.35	**84.02**	852	79
Sim_{Sup}	87.15	82.16	882	85
Sim_{ConEx}	89.27	83.49	856	82

Tabla 5.4: Resultados por medida de similitud

El valor de precisión más alto es del método de emparejamiento exacto, este método resultó ser el de mejores resultados debido a que existe un menor grado de equivocación al tener la restricción de aceptar sólo aquellas ENs que sean completamente iguales.

En cambio los otros métodos al tener un margen más libre en cuanto a las ENs que se consideran como similares y pertenecientes a la misma cadena se ven afectadas en su precisión cuando se aceptan elementos erróneos. En el caso de la métrica de coeficiente de superposición al estar basado en que se cumpla que un elemento sea subconjunto de otro, sin tomar en cuenta el orden de las palabras en los elementos obtiene los valores más bajos en términos de precisión.

5.2.3. Proceso de Validación de Resultados

Debido a que sólo se etiquetaron un conjunto de 250 documentos para la evaluación de las medidas de similitud, se creyó necesario realizar un proceso de validación para comprobar que los resultados obtenidos en el conjunto de documentos fuera consistente. Por esta razón se realizó un proceso de validación de resultados, donde se elige un número *n* de archivos de manera aleatoria y a este conjunto de archivos se le aplican los procesos de vinculación de ENs. Posteriormente se evalúan tanto precisión y recuerdo para los *n* archivos.

Para obtener el número *n* de archivos a evaluar se hace uso del muestreo. En estadística el tamaño de la muestra se refiere al número de sujetos que componen la *muestra* extraída de una población, necesarios para que los datos obtenidos sean representativos de la población. Esta selección de sujetos se realiza ya que en muchas ocasiones el tamaño de la población es muy grande y por eso, estudiarla completa resulta casi imposible. En muchas ocasiones, cuando se trata de seleccionar el tamaño de la muestra, se recurre a varios

criterios, básicamente empíricos y que muchas veces presentan subjetividad por parte de la persona que la selecciona, lo que hace que este aspecto carezca de un sustento matemático.

De ahí surge la necesidad de crear técnicas formales, que a partir de una pequeña, pero representativa muestra de la población, ayuden a inferir los resultados que se obtendrían si se analizara toda la población.

En la teoría del muestreo se deben considerar dos aspectos de suma importancia. Primero, el cálculo del tamaño de la muestra, este cálculo es importante ya que la muestra debe representar a toda la población y es por esto que se deben de considerar dos cosas, qué tan confiable es la muestra para generalizar resultados y también qué error puede existir en la muestra. El segundo aspecto a considerar es cómo seleccionar los elementos que serán parte de la muestra, en nuestro caso se usará el muestreo aleatorio simple, para que cada uno de los individuos de la población tenga la misma posibilidad de ser elegido, el muestreo aleatorio simple es un método sencillo y fácil de aplicar. Para calcular el tamaño de la muestra se deben de considerar los siguientes factores:

- El porcentaje de confianza con el cual se quiere generalizar desde la muestra hacia la población total.
- El porcentaje de error que se pretende aceptar al momento de hacer la generalización.
- El nivel de variabilidad que se calcula para comprobar la hipótesis.

La confianza o el porcentaje de confianza es el porcentaje de seguridad que existe para generalizar los resultados obtenidos. Esto quiere decir que un porcentaje del 100 % equivale a decir que no existe ninguna duda para generalizar tales resultados, pero también implica estudiar a la totalidad de los casos de la población.

Para evitar un costo muy alto para el estudio o debido a que en ocasiones llega a ser prácticamente imposible el estudio de todos los casos, entonces se busca un porcentaje de confianza menor. Comúnmente en las investigaciones se busca un 95 %.

El error o porcentaje de error equivale a elegir una probabilidad de aceptar una hipótesis que sea falsa como si fuera verdadera, o la inversa: rechazar la hipótesis verdadera por considerarla falsa. Al igual que en el caso de la confianza, si se quiere eliminar el riesgo del error y considerarlo como 0 %, entonces la muestra es del mismo tamaño que la población, por lo que conviene correr un cierto riesgo de equivocarse.

Comúnmente se aceptan entre el 4 % y el 6 % como error, tomando en cuenta de que no son complementarios la confianza y el error, en nuestro caso tomaremos un error de 5 %.

La variabilidad es la probabilidad (o porcentaje) con el que se aceptó y se rechazó la hipótesis que se quiere investigar en alguna investigación anterior o en un ensayo previo a la investigación actual. El porcentaje con que se aceptó tal hipótesis se denomina variabilidad positiva y se denota por p, y el porcentaje con el que se rechazó la hipótesis es la variabilidad negativa, denotada por q.

Hay que considerar que p y q son complementarios, es decir, que su suma es igual a la unidad: $p+q=1$. Además, cuando se habla de la máxima variabilidad, en el caso de no existir antecedentes sobre la investigación (no hay otras o no se pudo aplicar una prueba previa), entonces los valores de variabilidad son $p=q=0.5$.

Una vez que se han determinado estos factores, entonces se puede calcular el tamaño de la muestra como a continuación se expone. Como se conoce el tamaño de la población (que en nuestro caso son los 250 documentos), se

aplica la siguiente fórmula 5.11.

$$n = \frac{Z^2 pqN}{NE^2 + Z^2 pq} \tag{5.11}$$

Donde *n* es el tamaño de la muestra; *Z* es el nivel de confianza; *p* es la variabilidad positiva; *q* es la variabilidad negativa; *N* es el tamaño de la población y *E* es la precisión o el error.

Hay que considerar que debido a que la variabilidad y el error se pueden expresar por medio de porcentajes, hay que convertir todos esos valores a proporciones en el caso necesario.

También hay que tomar en cuenta que el nivel de confianza no es ni un porcentaje, ni la proporción que le correspondería, a pesar de que se expresa en términos de porcentajes. El nivel de confianza se obtiene a partir de la distribución normal estándar, pues la proporción correspondiente al porcentaje de confianza es el área simétrica bajo la curva normal que se toma como la confianza, y la intención es buscar el valor *Z* de la variable aleatoria que corresponda a tal área. Es decir si se quiere un porcentaje de confianza del *95 %*, entonces hay que considerar la proporción correspondiente, que es *0.95*. Lo que se buscaría en seguida es el valor *Z* para la variable aleatoria *z* tal que el área simétrica bajo la curva normal desde *-Z* hasta *Z* sea igual a *0.95*, es decir, *P(-Z<z<Z)=0.95*, así, basándose en tablas de distribución normal estándar, *Z=1.96*.

Utilizando la fórmula anterior tendríamos lo siguiente:

$$n = \frac{(1.96)^2(0.5)(0.5)(250)}{(250)(0.05)^2 + (1.96)^2(0.5)(0.5)} = 151.44 \tag{5.12}$$

Con lo cual se establece que el tamaño de muestra representativo serán 152 documentos; como ya se mencionó los documentos se eligen de forma aleatoria de la población total. Los resultados para las medidas de precisión y recuerdo se muestran en la Tabla 5.5.

Medida	**n=152**	
	P(%)	R(%)
Sim_{Ex}	***90.21***	*80.061*
Sim_{Dice}	*86.224*	***82.308***
Sim_{Sup}	*83.401*	*80.498*
Sim_{ConEx}	*84.81*	*81.359*

Tabla 5.5: Validación de resultados

Con estos resultados se puede observar que los valores de Precisión y Recuerdo son similares a los presentados en 5.4, es decir los valores presentados no varían drásticamente al verse reducido el número de archivos, el método de similitud exacta sigue obteniendo el valor de precisión más alto. Teniendo en cuenta estos resultados se decide utilizar en la segunda fase como método de similitud el método de similitud exacta de una cadena en otra, debido a que se prefiere que los elementos que van a formar parte de las cadenas sean elementos correctos y que pertenezcan a la cadena correcta ya que tomarán parte en el proceso de decisión de la nueva clasificación (ya sea dentro del enfoque de voto o como un atributo). Los resultados de la segunda etapa se verán influenciados por el resultado de la primera etapa, de ahí la importancia en elegir aquel método que brinde una mayor precisión.

Capítulo 6

Clasificación Utilizando Información Global

En este capítulo se aborda el último paso del método propuesto, que es la clasificación final de ENs. Como se ha mencionado anteriormente el objetivo de este segundo paso de clasificación es refinar una clasificación inicial. Se describen a continuación los distintos enfoques que se implementaron para realizar esta segunda clasificación. En primer lugar se describen los enfoques basados en voto, sección 6.1 y se presentan los resultados obtenidos, en la sección 6.1.3 se realiza un análisis de estos, para plantear distintas variantes buscando solventar las dificultades encontradas. Posteriormente, en la sección 6.2 se describe el enfoque utilizando árboles de decisión como método de clasificación, en esta sección se describen los atributos utilizados y en seguida las pruebas y resultados obtenidos. Finalmente se describe un análisis de los resultados y un análisis de los valores ideales.

6.1. Enfoques Basados en Voto

En distintas aplicaciones de aprendizaje automático se ha visto que no hay un solo algoritmo de clasificación que siempre funcione de forma adecuada para todos los casos, es por esto que han surgido distintos modelos que están compuestos de múltiples clasificadores que se complementan, por lo que se puede pensar que se pueden producir predicciones más confiables si se combinan las predicciones de varios modelos [1]. Así, a la combinación de varios modelos se les llama ensambles de clasificadores. Esta combinación se puede hacer de varias formas y una de las más sencillas es utilizar un sistema de voto, que corresponde a tomar una combinación lineal de distintos clasificadores. La idea de utilizar la información de las distintas apariciones de una EN es similar a la de los ensambles de clasificadores, sólo que en nuestro caso se aprovecha la información de la clasificación inicial de todas las ENs que estén vinculadas y se realiza una combinación de las clasificaciones iniciales con un esquema de voto.

En el segundo paso de refinamiento de clasificación que se definió en la arquitectura general se tienen los enfoques basados en voto los cuales tienen como objetivo considerar las clasificaciones iniciales de las ENs pertenecientes a una cadena de referencias e integrarlas en forma similar a un ensamble, de tal forma que estas clasificaciones tomen parte en la decisión de modificar la clasificación inicial obtenida a cada EN de la cadena. En un inicio se plantean dos métodos para alcanzar dicho objetivo, basados en considerar el número de apariciones de las clases en una cadena de ENs como votos. Los dos métodos son: el voto simple y voto ponderado, que se describen a continuación.

6.1.1. Voto Simple

En el primer método se considera un voto simple, el cual consiste en hacer una suma de las apariciones de cada clase en la cadena de ENs, es decir, sumar las veces que se asigna a las ENs de la cadena una determinada clase, la clase que sume un mayor número de votos será la clase que se asigne a todas las ENs pertenecientes a la cadena. Tomemos un ejemplo sencillo en la Figura 6.1, suponiendo que se tiene una cadena de entidades cuyas clases iniciales son LOC-LOC-ORG haciendo un voto simple se tendrían dos votos para la clase LOC y un voto para la clase ORG. Al tener LOC un mayor número de votos es la clase que se asigna finalmente a todas las ENs de la cadena, incluyendo aquella cuya clase inicial era ORG. A partir de aquí nos referiremos al voto simple como método 1.

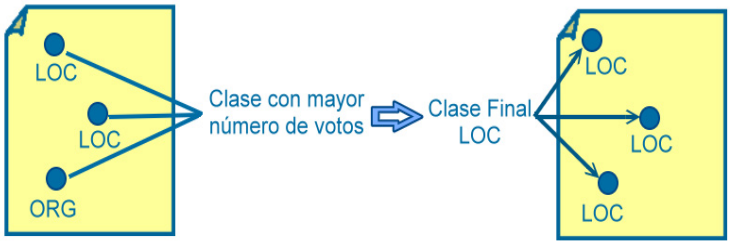

Figura 6.1: Ejemplo de voto simple

6.1.2. Voto Ponderado

Este método es una variante de la votación simple, donde además de la clasificación inicial se obtiene un valor de confianza asociado a la clase inicial. Este valor de confianza se obtiene de la herramienta FreeLing que asigna la clasificación inicial (como se describió en la sección 4.1). En lugar de realizar un voto simple se considera el valor de confianza y se van sumando los valores de confianza asociados a cada clase asignada a la EN. Por ejemplo (ver Figura

6.2), en el caso de que se tengan dos votos para la clase LOC y que sus valores de confianza sean 0.4 y 0.3, se obtiene la suma 0.7. En el caso de la clase ORG tiene un voto, con un valor de confianza de 0.8. Posteriormente se compara 0.7 de LOC con 0.8 de ORG y se asigna la clase ORG a todas las ENs de la cadena ya que tiene un mayor peso que la clase LOC. A partir de aquí nos referiremos al voto ponderado como método 2.

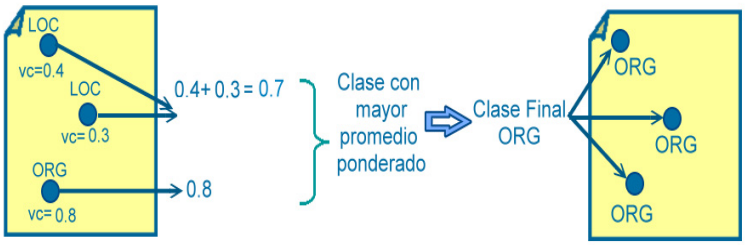

Figura 6.2: Ejemplo de voto ponderado

6.1.3. Resultados de los Enfoques Basados en Voto

Se realizaron un conjunto de pruebas con los corpus que se describen a continuación para los métodos de voto simple y voto ponderado.

6.1.3.1. Conjunto de Datos

- **Corpus CoNLL-2002 [7]**. Este corpus es una colección de cables de noticias de la Agencia EFE; los artículos que componen al corpus son de Mayo del año 2000. La anotación de los datos fue realizada por el Centro de Investigación TALP (Tecnologies i Aplicacions del Llenguatge i de la Parla) de la Universidad Técnica de Cataluña y el Centro de Lenguaje y Computación (CLiC) de la Universidad de Barcelona. Este corpus se encuentra dividido en tres partes que son: *esp.testa, esp.testb y esp.train*, donde el primer conjunto de datos se establece para la etapa de desa-

rrollo, *testb* para las pruebas y el último para entrenamiento. El conjunto de *entrenamiento* se encuentra constituido por 800 documentos, *testa* se compone de 157 documentos y *testb* de 200.

La clasificación inicial de los datos del corpus CoNLL para los propósitos de este trabajo se obtuvo bajo distintos esquemas, los cuales se describen a continuación.

- **Clasificación Inicial: FreeLing**. Una clasificación inicial es obtenida con la herramienta FreeLing, que es un conjunto de librerías para análisis del lenguaje que integra un módulo para el reconocimiento de entidades nombradas, este módulo utiliza una variante del algoritmo AdaBoost para la clasificación de ENs y fue desarrollada por Padró y Márquez (ver más detalles en [24]).

  El valor de confianza (v_{conf}) obtenido es la predicción de una variante del algoritmo AdaBoost. Cuanto mayor sea un valor positivo, la confianza de un objeto de pertenecer a una clase es más fuerte. Mientras más negativo es un valor, mayor confianza tiene el clasificador de que un objeto no pertenece a una clase. El algoritmo es una variante de AdaBoost en la cual las hipótesis pueden asignar confianzas a cada una de sus predicciones. Típicamente cada hipótesis débil es una regla que puede ser utilizada para generar una clasificación para cualquier instancia. En este caso cada hipótesis genera no solo clasificaciones predichas, sino también puntajes de confianza los cuales estiman la confiabilidad de cada una de sus predicciones. Para cada instancia x, una hipótesis débil h_t produce la predicción $h_t(x) \in \mathbb{R}$ cuyo signo es la etiqueta predicha (-1 o 1) y cuya magnitud $|h_t(x)|$ da una medida de "confianza" en la predicción. Sea $S = \langle (x_1, y_1), ..., (x_m, y_m) \rangle$ una secuencia de ejemplos de entrenamiento donde cada instancia x_i pertenece a un dominio o un

espacio de instancias X, y cada etiqueta y_i pertenece a un espacio de etiquetas finito Y. Se asume que se tiene acceso a un algoritmo de aprendizaje base o débil, el cual acepta como entrada una secuencia de ejemplos de entrenamiento S junto con una distribución D sobre $\{1, ..., m\}$, es decir, sobre los índices de S. Dada tal entrada, el aprendiz débil calcula una hipótesis débil h. En general, h tiene la forma $h : X \longrightarrow \mathbb{R}$. Se interpreta el signo de $h(x)$ como la etiqueta predicha (-1 o +1) a ser asignada a la instancia x y la magnitud $|h(x)|$ como la "'confianza" en esta predicción. Así, si $h(x)$ está cerca o lejos de cero, se interpreta como la confianza de predicción alta o baja. La versión utilizada difiere en que el rango de la hipótesis débil puede incluir todo $\mathbb{R}$ en lugar del rango restringido $\Big[-1,+1\Big]$. Esta medida de confianza es la que da como salida FreeLing y la que se utiliza como v_{conf} (ver más detalles en [26]).

- **Clasificación Inicial: Alicante.** El Laboratorio de Lenguaje Natural de la Universidad de Alicante proporcionó la clasificación inicial de los datos del *testb* del corpus CoNLL etiquetados bajo distintos esquemas: clasificación basada en máxima entropía y clasificación con el paquete de software TiMBL (aprendizaje basado en instancias con ganancia de información). Esta clasificación inicial se obtuvo con el trabajo de Kozareva et al. en [18]. Los números al lado del nombre de los datos indican el número de atributos utilizados, como se describe en [18]. Así para los etiquetados con MXE pertenecen a los datos de Máxima Entropía y varía el número de atributos, los etiquetados con TBL son aquellos clasificados con TiMBL, con variantes como la reducción de atributos (r), el uso de capitalización y caracteres (alfabéticos) dentro de los tokens (CIX) y el uso de la primera palabra de la entidad (fCIX). El método que utiliza el paque-

te TiMBL es un método basado en aprendizaje basado en memoria, trata un conjunto de instancias de entrenamiento representado como un espacio de características multi-dimensional, donde estas instancias son almacenadas en una especie de memoria. Posteriormente, las instancias de prueba se clasifican comparándolas con las instancias en memoria, mediante una función de distancia. En este caso se utiliza un algoritmo de k vecinos más cercanos para calcular la distancia entre las instancias que se encuentran en la memoria de instancias. El método de Máxima Entropía (ME) se basa en estimar probabilidades basándose en la idea de realizar un mínimo de hipótesis. La distribución de probabilidad que satisface esta propiedad será la que tenga la entropía más alta. Un clasificador que utiliza ME consiste en un conjunto de parámetros estimados mediante un proceso de optimización. Cada coeficiente está asociado con una característica de los datos de entrenamiento y el objetivo principal es obtener una distribución de probabilidad que maximice la entropía.

- **Corpus desastres.** Este conjunto de datos fue reunido y anotado por el Laboratorio de Tecnologías del Lenguaje del INAOE. Es un conjunto de noticias de desastres naturales de periódicos mexicanos en español. El número de documentos de este corpus es pequeño, ya que se tiene sólo una parte de los documentos etiquetados (con las etiquetas de ORG, LOC, ORG y PER) con respecto al corpus de desastres original, siendo 11 documentos con los que se realizan las pruebas. Este corpus no se divide en entrenamiento y prueba sino que se utiliza sólo como un conjunto de datos más para la evaluación de los métodos. La clasificación inicial se realizó con la herramienta FreeLing mencionada anteriormente, y por esto mismo se obtienen los valores de confianza asociados.

En la Tabla 6.1 se pueden observar algunos datos de los corpus CoNLL y Desastres. Cabe mencionar que para estos conjuntos de datos se tiene ya la vinculación de las ENs que las componen y se tiene la clasificación inicial obtenida bajo los distintos enfoques mencionados. Es importante resaltar que los conjuntos de datos cuya clasificación inicial fue obtenida por la herramienta FreeLing tienen además el atributo del valor de confianza (v_{conf}) con que es asignada una determinada clase a cada EN, en el caso de los datos de Alicante no se tiene este valor de confianza, por lo cual las pruebas que involucren este atributo no se realizan para estos datos.

C. Inicial	Datos	#Docs	#ENs	#Cadenas	#ENs en cadenas
CoNLL-FreeLing	Entrenamiento	*800*	*12,259*	*2,661*	*7,294*
	Testa	*203*	*7,612*	*575*	*1,573*
	Testb	*157*	*6,191*	*507*	*1,463*
	Desastres	*11*	*2,096*	*40*	*97*
CoNLL-Alicante	MXE_24	*157*	*6,089*	*490*	*1,372*
	MXE_25	*157*	*6,090*	*510*	*1,374*
	TBL_24	*157*	*6,089*	*490*	*1,372*
	TBL_24r	*157*	*6,087*	*490*	*1,374*
	TBL_24rCIX	*157*	*6,089*	*490*	*1,373*
	TBL_25fCIX	*157*	*6,090*	*490*	*1,374*

Tabla 6.1: Datos de corpus

6.1.3.2. Resultados

En la Tabla 6.2 se muestran los resultados de las pruebas para el etiquetado bajo FreeLing y el de Alicante, se incluyen los resultados para los datos de *entrenamiento* ya que los esquemas de votación son métodos no supervisa-

dos.

Se calcularon los valores de la medida F (descrita en la sección 2.3.1) para ambos métodos (voto simple y voto ponderado), en la primera columna de la tabla se muestran los valores de la medida F que se obtienen en el primer paso de clasificación, es decir, los valores de la clasificación inicial. De la misma forma y bajo los métodos de voto local y voto ponderado se realizaron pruebas con los conjuntos de datos de entrenamiento, testa, testb y para el corpus desastres, obteniendo los resultados mostrados en la Tabla 6.2; cabe mencionar que los datos mostrados se obtuvieron utilizando macro-promedio y es lo que se utiliza a los largo de las distintas pruebas.

C. Inicial	Datos	V.Iniciales (F %)	Met.1 (F %)	Met.2 (F %)
CoNLL-FreeLing	Entrenamiento	**69.14**	68.38	68.74
	Testa	**57.86**	57.19	57.81
	Testb	**68.71**	68.42	68.50
	Desastres	51.48	**54.53**	52.16
CoNLL-Alicante	MXE_24	62.98	**63.50**	-
	MXE_25	63.32	**64.30**	-
	TBL_24	61.68	**61.91**	-
	TBL_24r	62.79	**63.20**	-
	TBL_24rCIX	62.96	**63.34**	-
	TBL_25fCIX	61.27	**62.31**	-

Tabla 6.2: Resultados de métodos de voto

Como se puede observar en la Tabla 6.2 al aplicar el método de voto simple los resultados no sobrepasan los obtenidos por la clasificación inicial en el caso de los datos del corpus CoNLL bajo la clasificación inicial de FreeLing. Estos resultados se obtienen debido a la cantidad de etiquetas equivocadas de tipo

MISC (que podrían ser de tipo ORG, LOC o PER) que introducen ruido en varios casos, al hacer el voto mayoritario varias cadenas son clasificadas de este tipo. En el caso del corpus desastres y los datos de testb bajo la clasificación inicial de los métodos de Alicante, si se observa una mejora con respecto a la clasificación inicial con el método de voto simple, esto se debe a que en el caso de los datos de desastres el número de ENs clasificadas como Miscelánea es menor y por lo tanto involucran menos ruido en el paso de voto, es el mismo caso que con los datos de Alicante que existen menos votos erróneos de la clase Miscelánea.

En relación a la votación ponderada, se puede observar que los resultados son un poco más altos que en el caso del voto simple, sin embargo no presentan una mejora con respecto a la clasificación inicial. Si bien el valor de confianza asociado a la clasificación inicial es un aspecto que se consideró importante para apoyar la clasificación inicial, se presenta la dificultad de tener valores negativos en el caso de la clase MISC, esto quiere decir que esta clase siempre se encuentra en desventaja al realizar una suma ponderada, ya que independientemente del número de apariciones que presente en la cadena su valor será menor a aquellas entidades cuya clasificación inicial haya sido LOC, ORG o PER debido a que el valor de la suma de estas será positivo. Cabe recalcar que el voto ponderado no se realizó con los datos de Alicante ya que no se tiene asociado un valor de confianza a las clasificaciones iniciales de las entidades nombradas.

6.1.4. Análisis y Variantes de Enfoques Basados en Voto

Por las razones anteriormente expuestas, se realizó un análisis de los valores de confianza que se encuentran asociados a las ENs y sus clases asignadas, con este análisis esperamos obtener un método que nos permita corregir las ENs cuya clasificación inicial tenga un valor de confianza bajo, es decir,

que la clasificación inicial no sea completamente confiable, así como una manera de utilizar este valor de confianza sin afectar las ENs cuyo valor sea de negativo, esto es, sin penalizar las ENs de tipo miscelánea.

Este análisis se llevó a cabo para los datos de entrenamiento del corpus CoNLL al ser el conjunto de datos más grande. Algunos valores de este análisis se pueden ver en la Tabla 6.3, se lista la cantidad de ENs que tienen valores de confianza (v_{conf}) dentro de ciertos intervalos. En las columnas 2 y 3 se listan la cantidad de ENs que se encuentran en ciertos intervalos de valores de las ENs correctas, en la columna 2 el número de ENs cuyos valores de confianza se encuentran entre 0 y 1 y en la columna 3 las que sobrepasan el valor de 1. En las columnas 4 y 5 se encuentra la cantidad de ENs incorrectas y sus intervalos de confianza, en la columna 4 aquellas cuyos valores de confianza se encuentran entre 0 y 1, en la columna 5 aquellas mayores a 1. Cabe mencionar que sólo se muestran para las ENs clasificadas como ORG, LOC y PER, ya que las de tipo MISC tienen valores negativos y se describen posteriormente.

Clase	Correctas		Incorrectas	
	$0 < v_{conf} < 1$	$v_{conf} > 1$	$0 < v_{conf} < 1$	$v_{conf} > 1$
ENs	278	5350	123	163
ORG	137	2534	*57*	54
LOC	138	1373	56	*104*
PER	3	1463	10	5

Tabla 6.3: Análisis de valores de confianza

De la Tabla 6.3 se puede observar que el mayor número de ENs que son correctas tienen un valor de confianza mayor a 1, el número de ENs correctas menores a uno es pequeño en las 3 clases, en cuanto a las ENs incorrectas se ve que el número de incorrectas que sobrepasan el valor de uno es pequeño.

Los elementos incorrectos que son menores a uno son casi la mitad de los elementos incorrectos. Se observa también que los elementos incorrectos que sobrepasan el valor de 1 son en su mayoría los elementos de tipo LOC, estos valores nos dan bases para suponer que un valor de confianza de 1 puede utilizarse como umbral para modificar o no la clasificación inicial de una EN.

En la Tabla 6.4 se muestran los valores de confianza para la clase MISC, en las columnas 2 y 3 están la cantidad de ENs correctas; en la columna 2 las que tienen valores menores a -1 y en la columna 3 aquellas con valores entre -1 y 0. En las columnas 4 y 5 se encuentran la cantidad de ENs incorrectas, en la columna 4 aquellas con valores de confianza menores a -1 y en la columna 5 aquellas con valores entre -1 y 0.

Clase	Correctas		Incorrectas	
	$v_{conf} < -1$	$-1 < v_{conf} < 0$	$v_{conf} < -1$	$-1 < v_{conf} < 0$
MISC	1002	283	51	*133*

Tabla 6.4: Análisis de valores de confianza para MISC

En cuanto a la Tabla 6.4 se observa que la mayor parte de los elementos correctos son aquellos que tienen un valor de confianza menor a -1, son dos terceras partes de los elementos correctos, en cuanto a los elementos incorrectos la mayor parte se encuentra en este intervalo de 0 y -1, es decir no tienen un valor negativo tan marcado como aquellos que son correctos.

Realizando una tabulación de los elementos que se encuentran en el rango de 0 y -1 se encuentra que la mayoría son mayores a -0.5, es decir se van acercando a 0 y por lo tanto a los valores positivos. Se puede afirmar que muchas de las ENs incorrectas que se encuentran en el intervalo de 0 y -1 presentaban características que hacen dudar sobre si clasificarlas como ORG, LOC o PER al no alcanzar un valor positivo en estas clases y acercarse a los valores negativos se clasifican como MISC.

También del análisis se obtuvo que un número importante de ENs etiquetadas de tipo LOC son en realidad de tipo ORG, así como ENs clasificadas de tipo MISC corresponden realmente al tipo ORG o LOC, es decir las ENs de tipo PER no representan un número importante de errores, ni se produce confusión en su clasificación como en las otras tres clases.

Cabe mencionar que en estas cantidades no se incluyen aquellas ENs erróneamente clasificadas como ENs o cuya delimitación no coincide con la delimitación correcta, estas ENs erróneas fueron 793, donde la mayoría, 515, tienen un valor de confianza menor a 1 y sólo 237 tienen un valor mayor a uno.

En base a este análisis y los valores presentados en las tablas, se proponen distintos enfoques que tienen como objetivo aprovechar el valor de confianza obtenido pero sin penalizar aquellas ENs de tipo MISC, y se busca corregir aquellas ENs que tienen valores de confianza “débiles”, es decir cuya clasificación no fue completamente confiable y obtuvieron valores de confianza bajos. A continuación se listan los distintos enfoques propuestos, la mayoría de estos están basados en establecer un intervalo de valores, llamados $umbral_{mayor}$ y $umbral_{menor}$, para los valores de confianza (v_{conf}). Con umbrales definidos se busca probar que en estos intervalos se encuentran las ENs que tienen mayor probabilidad de ser erróneas y por lo tanto se definen los valores como un medio de decisión para cambiar o no la clasificación inicial de una EN. Con estos métodos se establecen los siguientes elementos: qué ENs son las que van a tomar parte del voto, qué ENs son las que se van a cambiar y porqué clase se va a realizar la modificación.

- Método 3. Este método se establece para comprobar qué tanto afecta el modificar sólo las ENs de clases ORG, LOC Y PER, y si inciden de manera positiva o negativa sin considerar la clase MISC. En este primer método se establece sólo un umbral, que es
  $umbral_{menor} = 1$, sólo se modifican las ENs que tengan un valor de con-

fianza asociado menor a 1; para este caso todas las ENs dan su voto. Las ENs de tipo MISC no tienen modificación alguna. La forma de establecer esta propiedad es la siguiente:

Sea $umbral_{mayor} = 1$. *Si* $v_{conf} < umbral_{menor}$ *entonces cambiar clasificación inicial de EN por voto mayoritario.*

- Método 4. Este método se propone para considerar la incidencia de las ENs con valor de confianza "fuerte" en el proceso del voto mayoritario, y si en su caso es mejor no considerar aquellas con confianza "débil". En este método se establecen dos umbrales. Para aquellas ENs que tengan un valor de confianza asociado que esté dentro del intervalo de 1 y -1 ($umbral_{mayor}\ y\ umbral_{menor}$) no se considera su voto para la clasificación final. Se modifican todas las ENs de la cadena por el voto mayoritario obtenido.

  Sea $umbral_{menor} = -1$ *y* $umbral_{mayor} = 1$. *Si* $umbral_{menor} < v_{conf} < umbral_{mayor}$ *entonces no sumar voto de EN.*

- Método 5. Este método propuesto pretende comprobar si no utilizar las ENs "débiles" permite obtener mejores resultados que cuando estos elementos forman parte del proceso de refinamiento de clasificación. En este método, se consideran los umbrales de 1 y -1 y se decide sólo modificar los elementos cuyo valor de confianza asociado se encuentre en ese intervalo, modificando también aquellas ENs que fueron clasificadas de tipo MISC a diferencia del primer método. En este caso todas las ENs participan en el voto mayoritario.

  Sea $umbral_{menor} = -1$ *y* $umbral_{mayor} = 1$. *Si* $umbral_{menor} < v_{conf} < umbral_{mayor}$ *entonces cambiar clasificación inicial de EN por voto mayoritario.*

- Método 6. Voto normalizado. Con el análisis realizado se observó que el

valor de confianza mostraba valores negativos y positivos, para utilizar estos valores en el mismo rango para las 4 clases, se normalizaron los valores de confianza para que estuvieran en un rango de 0 y 1, para determinar si utilizar este valor normalizado es de mayor utilidad. Este método es el mismo que el voto ponderado, donde se suman los valores de confianza, sólo que en este caso se suman los valores normalizados de los elementos. Se modifican todas las ENs de la cadena por el voto mayoritario.La fórmula para normalizar el valor de confianza se muestra a continuación 6.1.

$$V_{norm} = \frac{v_{conf} - min_{ai}}{max_{ai} - min_{ai}} \tag{6.1}$$

Dónde min_{ai} es el valor mínimo que pueden tomar los valores de confianza, y max_{ai} es el valor máximo que pueden alcanzar.

- Método 7. Este método propuesto pretende comprobar si no utilizar las ENs “débiles” en el proceso de voto y en el de modificación permitirá obtener mejores resultados que cuando estos elementos forman parte del de refinamiento de clasificación. En este método sólo se consideran los votos de aquellos elementos cuyos valores de confianza pasen los umbrales establecidos (de 1 y -1), la modificación de la clase se realiza bajo este mismo criterio, sólo se modifican las ENs que pasen el umbral. *Sea* $umbral_{menor} = -1$ *y* $umbral_{mayor} = 1$*.*

  Si $umbral_{menor} < v_{conf} < umbral_{mayor}$ *entonces no considerar voto de la EN y modificar clasificación inicial por voto mayoritario.*

- Método 8. Utilizar medida de distancia simple como método de decisión. Se propuso analizar si la distancia entre ENs podría servir como un factor a considerarse bajo la suposición de que generalmente cuando se habla de una misma EN la referencia a ella a lo largo del documento será la misma, como se describe en [30], un primer método fue calcular la distancia

en términos de oraciones entre ENs y asignar a la EN la clasificación de la EN más cercana, sólo en el caso de que la EN a modificar tenga un valor de decisión dentro del umbral definido.

Sea $umbral_{menor} = -1$ *y* $umbral_{mayor} = 1$.

Si $umbral_{menor} < v_{conf} < umbral_{mayor}$ *entonces cambiar clasificación inicial por clase de EN más cercana.*

- Método 9. En este caso se utilizó la distancia entre ENs pero de forma ponderada, utilizando el valor de confianza en relación con la distancia. Con la fórmula 6.2 se considera la distancia pero penalizando aquellas ENs que tengan un valor de confianza menor, donde d_o es la distancia a la que se encuentra la EN medida en oraciones.

$$dist_p = \frac{1}{d_o} * v_{conf} \tag{6.2}$$

Sea $umbral_{menor} = -1$ *y* $umbral_{mayor} = 1$.

Si $umbral_{menor} < v_{conf} < umbral_{mayor}$ *entonces cambiar clasificación inicial por clase de EN más cercana de forma ponderada.*

Los resultados para los enfoques mencionadas se muestran en la Tabla 6.5. Para estas pruebas se utilizaron los 250 documentos de los datos de *entrenamiento* del corpus ConLL con los que se evaluaron las cadenas en la primera fase, además de realizarse pruebas para los datos *testa, testb y desastres* (que fueron los corpus que tienen asociados valores de confianza). En los resultados, se puede observar que los mejores valores se obtuvieron empleando los umbrales de 1 y -1 (columna 5, método 3), sin embargo no ocurre esta situación en todos los casos, en el caso del corpus *testb* se obtiene cuando se consideran los umbrales para tomar en cuenta el voto de la EN y para modificar su clasificación (columna 7), pero sin lograr superar la clasificación

inicial. En el caso de los enfoques de distancia simple y ponderada se concluye que el involucrar esta medida como elemento para decidir el cambio de la clase no representa una mejora en la clasificación inicial. En este caso no se hicieron las pruebas con los datos de Alicante debido a que no se cuenta con el valor de confianza asociado a las clasificaciones iniciales.

C. Inicial	Datos	Resultados por método							
		V.Iniciales	Met.3	Met.4	Met.5	Met.6	Met.7	Met.8	Met.9
CoNLL-FreeLing	Entrenamiento	71.5	71.16	71.67	**71.68**	70.74	71.66	70.58	71.37
	testa	68.71	68.46	67.27	68.58	68.35	**68.63**	67.64	68.54
	testb	57.86	57.93	56.67	**58.06**	57.84	58.01	57.52	57.97
	Desastres	51.48	**53.51**	49.91	53.15	50.96	49.59	52.81	52.81

Tabla 6.5: Resultados de variantes de voto

6.1.5. Análisis de Distribución de Clases en Cadenas

Después de realizar el análisis de los resultados con los enfoques de voto y sus distintas variantes, se propuso realizar un análisis de la ocurrencia de las distintas clases en las cadenas que se forman en el primer paso, analizar si la mayoría está compuesta de una clase, 2, 3 o incluso las 4 clases. Se realiza este análisis para comprobar si efectivamente las cadenas originales están compuestas tan sólo por una clase y para comprobar la distribución de las clases en las cadenas, esta distribución nos permitirá analizar que clases introducen más ruido y si es posible corregir estos errores distribuyendo las clases de manera diferente. Con este análisis se espera conseguir un conjunto de métodos que permita corregir la aparición de clases incorrectas en las cadenas, reduciendo la aparición de clases por cadenas.

En el paso de vinculación de ENs, al enlazar las ENs se obtienen cadenas que en muchos casos no tienen una sola clase, muchas de estas cadenas se componen de 2 o 3 clases al etiquetarse las ENs de distintos tipos, en la Tabla 6.6 se puede observar la distribución de estas.

En la Fila encabezada como *Entrenamiento FreeLing* se listan los resultados obtenidos por la clasificación inicial de FreeLing y con el método de vinculación explicado en el capítulo anterior. En la Fila encabezada como *Original* se listan los resultados que se obtienen al utilizar la clasificación original (la clasificación correcta de las ENs) y al realizar el método de vinculación para enlazar las referencias de las ENs.

En las columnas 2-5 encabezadas por *#clases por cadena* se encuentra el número de clases que se obtienen en las cadenas de salida, es decir, si una cadena está compuesta por elementos de una sola clase o puede tener elementos de dos clases, aún cuando las ENs se escriban de la misma manera. Se puede apreciar las diferencias cuando el número de clases son 2, 3 y 4, que en el caso de los datos originales son mucho menores que con FreeLing, es decir, predominan las cadenas que son de una sola clase.

En las columnas encabezadas como *1 clase* se lista el número de cadenas por clase que se obtienen, en este caso para las clases ORG y LOC la diferencia entre números es de 100 cadenas. En las columnas encabezadas como *2 clases* se lista la forma en que están compuestas las cadenas de 2 clases, podemos ver que el número de elementos es mucho menor en los datos originales, y que la distribución se centra en cadenas de tipo ORG-LOC, sin embargo, con los datos de FreeLing se concentran también en cadenas que tienen clase MISC como una de las dos clases involucradas. En las columnas encabezadas como *3 clases* se listan el número de cadenas que involucran 3 clases, en este caso se puede apreciar nuevamente que el número de veces que esto ocurre en el etiquetado correcto es mínimo, teniendo sólo 2 casos de este tipo, sin embargo en FreeLing se tienen 28 cadenas con entidades de 3

clases. Finalmente, en la última columna se listan las cadenas que tienen ENs de 4 clases distintas en una misma cadena, para el etiquetado original estos casos son nulos, pero bajo FreeLing ocurren 2 casos.

El número de cadenas que se obtienen con la vinculación de las ENs de la clasificación inicial es de 2626 y el número de cadenas original es de 2661. Podemos darnos cuenta de que el número de cadenas obtenidas no varía demasiado, lo que varía es la distribución de las clases en las cadenas. En las cadenas que se forman con la clasificación inicial existen cadenas de 2, 3 e incluso 4 clases, mientras que en el etiquetado original predominan las cadenas de 1 y 2 clases sin existir aquellas de 3 o 4 clases. Se puede notar la desigualdad en la distribución por ejemplo en las cadenas de tipo ORG-MISC que en el etiquetado original son 19 y en el etiquetado inicial es de 100 cadenas, lo mismo se observa en las cadenas LOC-MISC donde la relación es de 61 con el etiquetado de FreeLing y de 5 en las cadenas originales. Con estos datos se puede observar que las ENs que son clasificadas de tipo MISC son las que parecen presentar “ruido” en las cadenas, sin embargo es difícil decidir si cambiar o no estas ENs y porqué clase deberían se cambiadas.

	# clases por cadena				1 clase				2 clases						3 clases			4 clases			
	1	2	3	4	ORG	LOC	PER	MIS	ORG LOC	ORG MIS	ORG PER	LOC MIS	LOC PER	PER MIS	ORG LOC PER	ORG LOC MIS		ORG	LOC	PER	MIS
Entren. FreeL-ing	2288	308	28	2	906	520	578	284	123	100	1	61	6	17	0	28		2			
Original	2487	172	2	0	1048	601	590	248	137	19	5	5	4	2	1	1		0			

Tabla 6.6: Distribución de clases por cadenas

En base al análisis de las cadenas y a la observación de la distribución de las clases, se proponen 4 métodos basados en la modificación del número de clases por cadena, eliminando principalmente una clase “débil” para favorecer la aparición de la clase mayoritaria (clase fuerte) y conservar una distribución similar a la que se presenta en los datos originales. Las pruebas que se realizaron siguiendo este enfoque se listan a continuación, dónde *clases_cadena* se refiere al número de clases distintas por cadena, *clase_minoritaria* es la clase que aparece menor número de veces en la cadena y *clase_mayoritaria* la que aparece con mayor frecuencia:

- Método 10. Este método se propone con la idea de favorecer la clase fuerte en la cadena, para analizar si el eliminar la clase “débil” contribuye de manera positiva en la clasificación. Cambia clase minoritaria por clase mayoritaria convirtiendo cadenas de 3 clases en 2 y de 2 clases en 1. Esto se hizo eliminando la clase más “débil” (basándose en los valores de confianza de las ENs) y dejando la clase fuerte en su lugar. En este caso todas las ENs participan en el voto mayoritario.

  Si clases_cadena=3 o clases_cadena=2 entonces hacer voto mayoritario y cambiar sólo clase_minoritaria por clase_mayoritaria.

- Método 11. Cambia clase minoritaria por clase mayoritaria, sólo cadenas que no sean de tipo ORG-LOC, es decir cadenas que sean de tipo ORG-MISC, LOC-MISC, PER-MISC, ORG-PER, LOC-PER, ORG-LOC-MISC serán modificadas, debido a que se observa que la clase MISC es la que presenta mayores dificultades.

  Si clases_cadena=2 y no son del tipo ORG-LOC entonces hacer voto mayoritario y cambiar sólo clase_minoritaria por clase_mayoritaria.

- Método 12. Si la clase minoritaria es de tipo MISC se cambia por ORG o LOC, en el caso de que la clase minoritaria sea de otro tipo la cadena

permanece igual, realizar los cambios indicados se basa en el hecho de que se observa que la clase MISC presenta mayor ruido y errores en la cadena de ENs, se creé que eliminando estos elementos erróneos los resultados serán mejores.

Si clase_minoritaria=MISC entonces cambiar por ORG si es la clase_mayoritaria o por LOC si es la clase_mayoritaria.

- Método 13. Este método se propone basándose en los resultados del método 3, para comprobar la incidencia de las ENs débiles en la distribución de las cadenas. Además de que se busca disminuir el número de cadenas con 3 clases, ya que se observa que en el corpus original estas cadenas no se presentan. Sólo se cambian cadenas que tengan ENs con 3 clases diferentes, asigna clase mayoritaria a ENs que no pasen el umbral de 1 y -1. En este caso todas las ENs participan en el voto mayoritario.

  Si clases_cadena=3 y $-1 < v_{conf} < 1$ *entonces cambiar EN por clase_mayoritaria.*

Las pruebas de estos enfoques se realizaron con los documentos de los corpus de *entrenamiento, testa, testb, y desastres* bajo la clasificación de FreeLing. Los resultados de estas pruebas se muestran en la Tabla 6.7, nuevamente estas pruebas no se realizaron con los datos de Alicante debido a que no cuenta con los valores de confianza asociados a las ENs.

Al igual que en las pruebas anteriores en la segunda columna se muestra el valor de la clasificación inicial. Se puede ver que los mejores valores se consiguen con el método 13, ya que en todos los corpus se tiene un mayor valor al obtenido por la clasificación inicial. Estos resultados indican que la existencia de cadenas de 3 clases es mínima y en la mayoría de los casos incorrecta, por lo que realizar un cambio por la clase más “frecuente” pudiera ser una buena opción cuando se tienen pocos elementos de la clase minoritaria. Otro punto

C. Inicial	Datos	Resultados				
		V.Iniciales	Met.10	Met.11	Met.12	Met.13
CoNLL-FreeLing	Entrenamiento	71.5	70.94	71.09	71.07	**71.55**
	testa	68.71	68.39	68.68	68.34	**68.74**
	testb	57.86	57.82	57.75	57.53	**58.05**
	desastres	51.48	53.82	53.84	53.82	**54.79**

Tabla 6.7: Resultados de enfoques de cadenas

que cabe resaltar es la utilidad del umbral para hacer la modificación, en este caso se cambian las ENs que no pase el umbral corrigiendo aquellas que no tengan una seguridad fuerte en su clasificación inicial. Se puede apreciar que en el caso del Corpus desastres se obtienen buenos resultados, creemos que esto se debe a que el dominio de este corpus es más reducido, aún cuando se tratan de noticias (como en el caso del corpus CoNLL) estás noticias tienen un dominio más específico, que sólo se centran en noticias de desastres naturales, donde la mayoría de las cadenas de ENs tienden a ser de una sola clase, y es por esta razón que los métodos propuestos que buscan corregir la presencia de otras clases en una cadena para dejar una sola clase por cadena funcionan de manera favorable. Además el número de documentos es otro elemento relevante, se tienen pocos documentos y pocas ENs, cuando se corrigen correctamente esto tiene un impacto más notable.

6.2. Enfoques basados en Árboles de Decisión

Después de plantear distintos enfoques basados principalmente en el voto de las clasificaciones iniciales de las ENs se presenta el segundo enfoque

de clasificación, basado en árboles de decisión. Se decidió hacer la transición de los métodos anteriores a un método supervisado, principalmente pensando en la forma de aprovechar la información de una primera clasificación como un atributo y aprender de los errores de esa clasificación inicial. Los métodos anteriores utilizan la información de manera más informal, basados en las observaciones y análisis de los datos a diferencia de un método supervisado que “aprenderá” a generalizar las características que permitan distinguir entre clases, además se busca aprovechar e integrar la clase inicial como un atributo más en el árbol de decisión.

6.2.1. La Información Global como Atributos

El objetivo de integrar las ENs en un conjunto de cadenas es aprovechar la información de las ENs que pertenecen a ellas y que dicha información pueda apoyar su clasificación. Para alcanzar este objetivo se consideró integrar la información de las cadenas como atributos para realizar una nueva clasificación. Por esta razón se eligieron un conjunto de atributos que integran información de las ENs y de las cadenas producto de la vinculación de ENs, el primer atributo es una propiedad de la EN a clasificar y los siguientes 4 son atributos que se obtienen de la cadena en la que se encuentra la EN, los atributos se listan a continuación:

- Valor de confianza: El valor de confianza asociado a la clase que se asignó a una EN, este valor se utilizó sin normalizar.
- Clases en cadena: Este atributo indica otras clases que estén presentes en las ENs de la cadena, los valores que puede presentar el atributo son: *O, L, P, M, OL, OP, OM, LP, LM, PM, OLP, OLM, LPM, OPM* y por último *none* en el caso de que no exista otra clase en la cadena.
- Número de ocurrencias: El número de veces que aparece en la cadena

la clase de la EN que se va a clasificar.

- Clasificación inicial de vecino izquierdo y derecho: la clasificación inicial de la ENs que le antecede y que le sucede en la cadena a la EN a clasificar.

- num_org, num_loc, num_per y num_misc: El número de veces que aparecen otras clases en la cadena de la EN a clasificar.

En el caso del atributo valor de confianza, se entrenó un clasificador sin tomar en cuenta este atributo, sólo considerando los atributos anteriormente listados, y un clasificador considerando este valor junto con los demás atributos, esto para comprobar la inferencia del valor de confianza en la clasificación de las ENs.

Para construir los árboles de decisión se busca generar un conjunto de reglas que permitan corregir errores de la primera clasificación, por esta razón, se construyen árboles para generar reglas para cada una de las clases MISC, ORG, LOC y PER. cada árbol se construye utilizando como instancias de entrenamiento aquellas instancias cuya clasificación inicial sea de la clase para la que se va a construir el árbol, por ejemplo, si se va a construir el árbol para generar las reglas para corregir la clase ORG, se construye con las instancias cuya clase inicial fue ORG y se incluye la clase verdadera para que el árbol pueda corregir los errores, es decir, para aprender si se debe de mantener la clase inicial (en este caso ORG) o se debe cambiar por LOC, PER O MISC. Se construyen los árboles de esta manera porque se quiere aprender de forma individual los cambios que deben realizarse para la clase inicial.

Los árboles de decisión para las clases de miscelánea, organización y persona se muestran en las Figuras 6.3, 6.4 y 6.5. Para la clase LOC se obtiene un árbol de decisión de una sola hoja, todos los elementos que se encuentran de tipo LOC como clasificación inicial deben dejarse con esa clase. Estos

árboles de decisión son los construidos con los atributos mencionados anteriormente, donde se utiliza el valor de confianza.

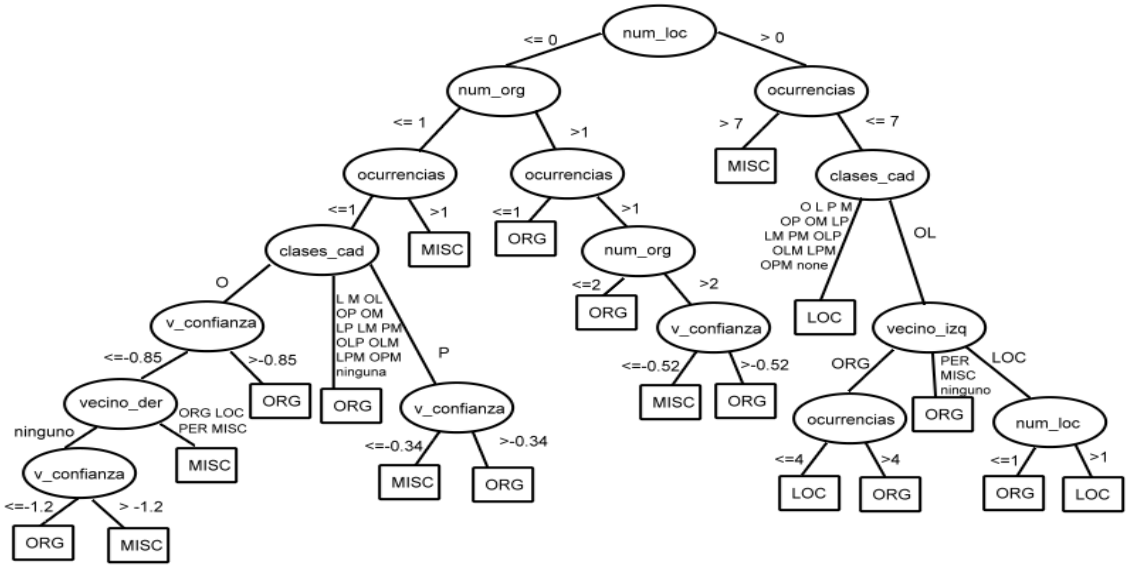

Figura 6.3: Árbol de decisión con valor confianza MISC

De los árboles obtenidos se puede observar que el que presenta mayor complejidad es el árbol de la clase Miscelánea, lo cual es un reflejo de la dificultad para clasificar los elementos en esta clase. De este árbol se obtienen reglas del siguiente tipo:

Si num_loc >0 y ocurrencias ≤ 7 *y clases_cad* $=$ *OL y vecino_izq* $=$ *ORG y ocurrencias >4 entonces clase es ORG*

Con esta regla se cubren los casos donde las cadenas están compuestas por 3 clases, miscelánea, organización y lugar, se toman en cuenta el número de veces que aparece la clase miscelánea y si el vecino a la antecesor es de tipo organización, entonces se asigna organización como la nueva clase. Posteriormente se agregaron otros atributos que se consideró podrían apoyar la clasificación, tomando en cuenta atributos de elementos vecinos que rodean a la EN además de otros atributos de las ENs vecinas en la cadena. Los atributos agregados son el elemento PoS, la clasificación inicial (CEN) de los elementos

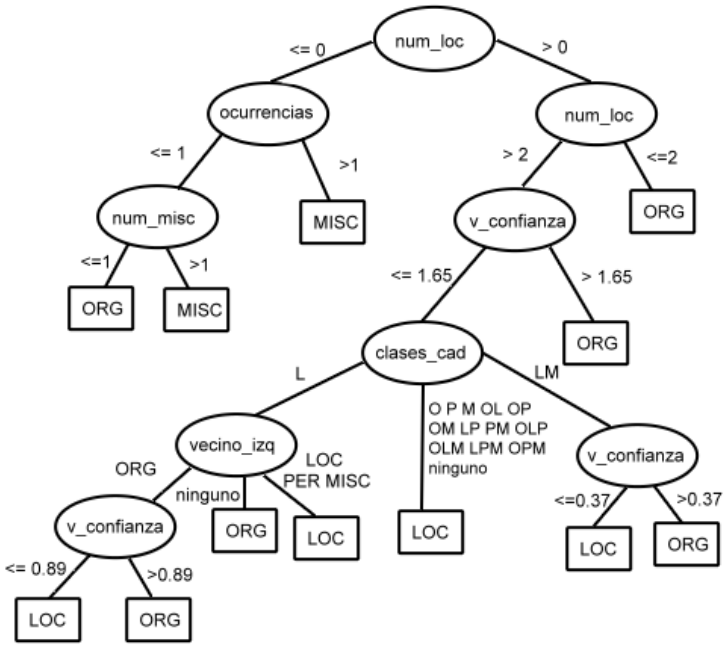

Figura 6.4: Árbol de decisión con valor confianza ORG

vecinos y es_acronimo que permite identificar si existen acrónimos expandidos (que se expandieron en el preprocesamiento) en otros elementos de la cadena, lo cual puede apoyar las cadenas de tipo organización. La totalidad de atributos se lista a continuación:

- es_acronimo: Un valor booleano que indica si la EN a clasificar es un acrónimo de alguna otra que se encuentra en la cadena.
- POS_elemento: El valor de etiqueta POS(Part of Speech) o parte de la oración, que se obtuvo del etiquetador en la primera fase de clasificación final.
- POS_izquierdo y POS_derecho: Se agregan también los valores de la etiqueta POS de elementos vecinos izquierdo y derecho que rodean a la EN.

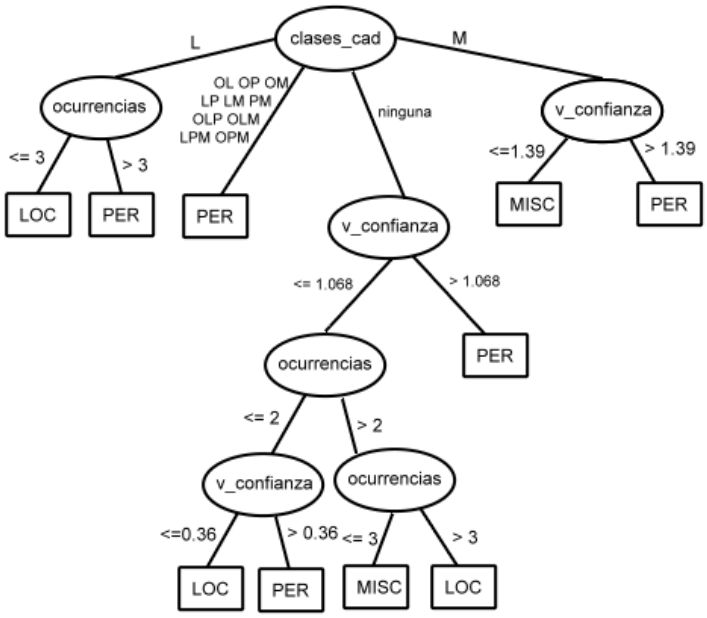

Figura 6.5: Árbol de decisión con valor confianza PER

- CEN_izquierdo y CEN_derecho: Se agregan los valores de clasificación inicial de elementos vecinos izquierdo y derecho, que pueden tener las etiquetas de NP00S00, NP00O00, NP00G00, NP00SP0 y O, esta última etiqueta indicando que el elemento no es una EN.

- clases en cadena: Este atributo indica otras clases que estén presentes en las ENs de la cadena, los valores que puede presentar el atributo son: *O, L, P, M, OL, OP, OM, LP, LM, PM, OLP, OLM, LPM, OPM* y por último *none* en el caso de que no exista otra clase en la cadena.

- número de ocurrencias: El número de veces que aparece en la cadena la clase de la EN que se va a clasificar.

- POS, CEN y es_acronimo de la EN que le antecede y de la EN que le sucede; Se agregan también los atributos de ENs que le preceden y que le suceden considerando la posición de la EN en la cadena,

- num_org, num_loc, num_per, num_misc: El número de veces que aparecen otras clases en la cadena de la EN a clasificar.

Se entrena un árbol de decisión para cada clase utilizando el conjunto de atributos de la primera lista (con y sin valor de confianza) y el conjunto de atributos de la segunda lista. Se obtiene entonces un árbol de decisión por clase. Los árboles obtenidos utilizando los atributos vecinos se muestran en las Figuras 6.6, 6.7 y 6.8 para las clases miscelánea, organización y persona respectivamente, en este caso nuevamente las ENs de clase inicial LOC se mantiene sin cambio.

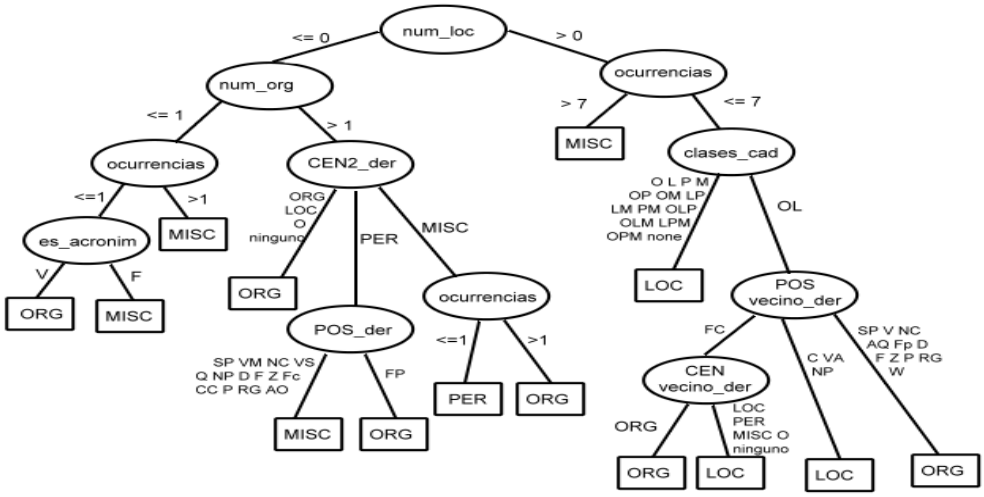

Figura 6.6: Árbol de decisión atributos vecinos: MISC

Cabe mencionar que en cuanto a los elementos vecinos, se realizaron pruebas para determinar el número de elementos vecinos a la izquierda y a la derecha de la EN a considerar, pruebas con ventanas de tamaño 1, 2 y 3 elementos a la izquierda y a la derecha, pero se determinó que tomar un solo elemento antecesor y sucesor eran suficientes, ya que tomando una ventana de 2 y 3 elementos vecinos los resultados de clasificación obtenían resultados menores, es decir, afectaban el proceso de clasificación de manera negativa.

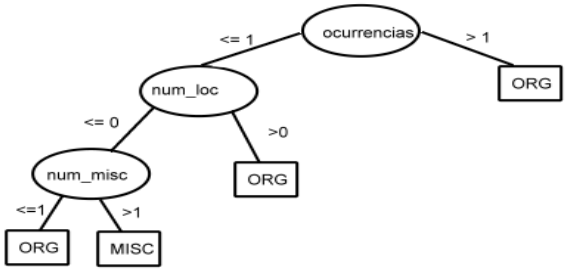

Figura 6.7: Árbol de decisión atributos vecinos: ORG

Figura 6.8: Árbol de decisión atributos vecinos: PER

6.2.2. Resultados de Árboles de Decisión

Los árboles de decisión se entrenaron con los 800 archivos de los datos de *entrenamiento* utilizando el enfoque de validación cruzada con 10 pliegues. Se realizaron distintas pruebas con los datos de *testa, testb, desastres* y los datos de la Universidad de Alicante, *MXE24, MXE25, TBL24, TBL24r, TBL24rCIX, TBL25fCIX*. Los resultados de estas pruebas se listan en la Tabla 6.8. El método 14 se refiere a los árboles de decisión que no integran el valor de confianza como atributo, el método 15 si lo incluye y el método 16 los árboles de decisión que utilizan los atributos vecinos.

En la segunda columna de la tabla se tienen los valores obtenidos por la clasificación inicial, en la tercera columna los obtenidos por la primer lista de atributos sin tomar en cuenta el valor de confianza, en la cuarta columna los porcentajes que se obtienen considerando el valor de confianza y en la últi-

C. Inicial	Datos	Resultados			
		V.Ini-ciales	Met.14	Met.15	Met.16
CoNLL-FreeLing	testb	68.71	**68.80**	68.73	68.735
	testa	57.86	57.84	57.70	**57.94**
	desastres	51.48	**55.25**	50.40	54.29
CoNLL-Alicante	MXE24	62.98	62.97	-	**63.01**
	MXE25	63.32	63.28	-	63.30
	TBL24	61.68	**61.74**	-	**61.74**
	TBL24r	62.79	62.83	-	**62.90**
	TBL24rCIX	62.96	63.03	-	**63.05**
	TBL25fCIX	61.27	**61.94**	-	61.82

Tabla 6.8: Resultados de árboles de decisión

ma columna se tienen los resultados integrando información de los elementos vecinos. En el caso de los datos de entrenamiento, testa, testb y desastres se puede observar que el valor de confianza como un atributo más no proporciona una mejora en la clasificación, en el caso del corpus de desastres incluso la empeora, por lo que se concluye que el valor de confianza no se puede tomar como un valor decisivo para cambiar la clasificación. Se puede ver que en la mayoría de los casos tomar los atributos de los elementos vecinos puede impactar de forma positiva en la clasificación, tan sólo testb y TBL25fCIX obtienen mejores resultados sin utilizar los atributos vecinos, en los demás corpus si se obtiene una mejora en la clasificación.

6.3. Voto entre Documentos

Como se ha mencionado a lo largo de este trabajo se hace uso de la búsqueda de referencias de ENs en un documento, es importante mencionar que no todas las ENs se encontrarán encadenadas, las ENs que sólo se mencionan una vez en un documento les llamaremos ENs “sueltas”, es decir son aquellas ENs que no pertenecen a ninguna cadena. Estas ENs con los métodos anteriores permanecen sin cambio y no se involucran de forma alguna en los métodos de decisión de modificar las clases de las ENs en cadenas ya que no son referencia de ninguna otra EN en el documento que se está revisando. Sin embargo, a continuación se propone utilizar estás ENs “sueltas” como un elemento más de información. A continuación, se propone un último acercamiento al uso de información global para la modificación de la clase inicial de ENs, involucrando no sólo la información de un sólo documento, sino de varios documentos donde se encuentren las ENs. Con esto, se busca analizar el impacto que pudiera tener el voto de ENs “sueltas”, es decir, si podrían aportar mayor información o información más confiable que la que se aporta por las ENs encadenadas.

Se consideró como último método tomar en cuenta un voto entre documentos, considerando no sólo las ENs que se encuentran en cadenas sino también aquellas ENs que no pudieron encadenarse en un documento pero que podrían servir de referencia en otros documentos. Basados en esta idea se propusieron dos enfoques:

- **Método 17.** Considerar el voto de ENs sueltas y las ENs en documentos. En este caso se enlazan las apariciones de las ENs en un documento y con aquellas que se encuentren “sueltas” (que no pertenezcan a una cadena) en otros documentos, es decir se enlazan todas las apariciones de las ENs en todos los documentos que se encuentren (ver Figura 6.9),

en este caso se asigna la clase mayoritaria a todos los elementos que pertenezcan a la misma cadena.

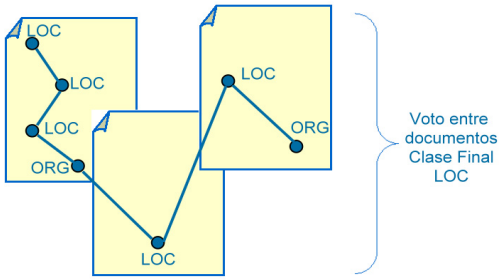

Figura 6.9: Voto entre documentos. Método 17

- **Método 18.** Un segundo método fue considerar a las ENs de una cadena como un solo voto, es decir en un primer paso se realiza el voto simple entre las cadenas de un documento, obteniendo un voto por cadena y después se enlazan con las EN “sueltas” en otros documentos (ver Figura 6.10), una vez que se enlazan con las ENs “sueltas” se realiza otro voto, asignando la etiqueta mayoritaria a todas las ENs.

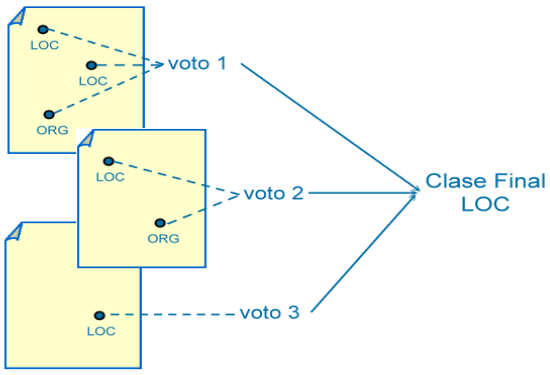

Figura 6.10: Voto entre documentos. Método 18

Los resultados de estos métodos se muestran en la Tabla 6.9.

C. Inicial	Datos	Resultados entre documentos			
		V.Ini-ciales	Met.1	Met.17	Met.18
CoNLL-FreeLing	Entrenamiento	69.14	68.38	68.82	68.17
	testb	68.71	68.42	68.18	68.23
	testa	57.86	57.19	57.76	57.76
	desastres	51.48	54.53	54.00	**54.55**
CoNLL Alicante	MXE_24	62.98	63.50	**63.54**	63.43
	MXE_25	63.32	**64.30**	64.05	63.75
	TBL_24	61.68	**61.91**	62.01	61.90
	TBL_24r	62.79	63.20	**63.23**	62.81
	TBL_24rCIX	62.96	63.34	**63.46**	62.88
	TBL_25fCIX	61.27	62.31	62.28	**62.33**

Tabla 6.9: Resultados de voto entre documentos

En la primera columna se encuentran los datos de la clasificación inicial para cada uno de los corpus. En los resultados obtenidos se creyó importante incluir los resultados de voto local (tercera columna, método 1) reportados en la primera sección para comparar el impacto de votos de otros documentos en la clasificación, en la cuarta columna se incluyen los resultados del método 17 y en la quinta columna los del método 18. Se puede observar que el método 18 (columna 5) obtuvo mejores resultados para los datos de Alicante en la mayoría de los casos. El método 17 obtiene mejores resultados para los datos de Alicante en todos los casos con respecto a los valores iniciales, lo cual se basa en obtener buenos resultados en el primer paso de un voto simple, sin embargo este método no funciona para el caso de los datos de *entrenamiento, testa, testb y desastres* (con FreeLing) al no obtener buenos resultados en el primer caso de voto simple, esto repercute en que el voto que se da por cadena

puede no ser correcto, afectando la clasificación final.

6.4. Análisis de Valores Ideales

En vista de los resultados obtenidos se creyó necesario realizar un análisis de los valores y de la clasificación final obtenida. Se calcularon los valores de Precisión, Recuerdo y Medida F que se obtienen con la clasificación inicial de FreeLing y se midieron los valores que se alcanzarían si se cambiaran las clases incorrectas de las ENs que pertenecen a una cadena por las clases correctas. A estos valores se les llaman valores ideales, debido a que son los valores que se alcanzarían si tuviéramos una clasificación perfecta en las ENs encadenadas, es decir una clasificación ideal.

En la Tabla 6.10 se exponen algunos valores para analizar los resultados obtenidos. En la tercera columna se muestran los valores obtenidos bajo la clasificación inicial, en la cuarta columna encabezada como *Ideal*, se muestran los valores en términos de la medida F, estos valores representan el porcentaje que se pudiera alcanzar en el caso de que se corrigieran toda las ENs erróneas que se encuentran en cadenas, como podemos ver, el aumento en porcentajes en la mayoría de los casos es pequeño con respecto a los valores iniciales. En la quinta columna se lista la diferencia entre el valor inicial y el valor ideal, esto para obtener cuánto es lo máximo que podría mejorar la clasificación final, encontrándose estos valores entre 1.64 y 4.98; en la sexta columna se encuentran los resultados relativos con voto local (Método 1), en la séptima columna los resultados relativos obtenidos con los árboles de decisión utilizando atributos vecinos (Método 16) y en la última columna los resultados con voto entre documentos (Método 17), se listan estos resultados debido a que fueron aquellos donde los conjuntos de datos presentaron mejores resultados. Para obtener los resultados relativos se realiza una resta entre el valor de la clasificación final que obtiene el método y el valor de la clasificación ini-

cial, el resultado de esta resta es el incremento o decremento que se alcanza con respecto a la clasificación inicial, este valor es entonces comparado con el incremento ideal que se podría alcanzar (utilizando porcentajes), para obtener un resultado relativo al valor ideal, es decir, qué tanto se incrementó la clasificación final con respecto al mayor valor que podría alcanzarse.

C. Inicial	Datos	Resultados relativos					
		V.Iniciales	ideal	Diferencia	Met.1	Met.16	Met.17
CoNLL-FreeLing	Entrenamiento	69.14	72.19	3.04	-25 %	**8.38 %**	-31 %
	testb	68.71	71.87	3.16	-9.71 %	**0.072 %**	-15 %
	testa	57.86	59.5	1.64	-40 %	**4.86** %	-6.09 %
	desastres	51.48	55.70	4.22	72.27 %	66.58 %	**72.74 %**
CoNLL-Alicante	MXE_24	62.98	66.99	4.003	12.99 %	0.5 %	**13 %**
	MXE_25	63.32	67.31	3.98	**24.79 %**	-5 %	18 %
	TBL_24	61.68	65.96	4.28	5.37 %	1.4 %	**7.7** %
	TBL_24r	62.79	66.91	4.12	9.95 %	2.66 %	**10.6** %
	TBL_24rCIX	62.96	67.17	4.21	9.06 %	2.13 %	**11.8** %
	TBL_25fCIX	61.27	66.26	4.98	**20.88** %	11 %	20.18 %

Tabla 6.10: Resultados de valores ideales

Hay que tomar en cuenta que los valores ideales a alcanzarse presentan un margen de mejora pequeño, ya que el número de elementos en las cadenas incorrectos no influyen de manera significativa en el porcentaje general de los datos. En el caso de los valores negativos se calculan de igual manera, tomando en cuenta la proporción en que disminuye el valor de la medida F con respecto a la diferencia entre el valor ideal y el valor inicial, es por esto que estos valores son mayores.

Los resultados entre los datos de *Alicante* y los obtenidos para los datos de

entrenamiento, testa, testb y desastres de FreeLing varían entre los métodos propuestos debido a que los errores que existen entre aquellos etiquetados por FreeLing y los etiquetados bajo los métodos de Alicante no son los mismos. En el caso de los etiquetados por FreeLing el principal problema es la clase MISC que dificulta la asignación correcta de la segunda clasificación, en el caso de los datos de Alicante los errores se encuentran más distribuidos entre las 4 clases y es por esta razón que no se penaliza de alguna manera a alguna clase específica.

Otro problema son los elementos mal delimitados en la primera fase, estos elementos aún cuando se encuentren en cadenas son difíciles de corregir, ya que la delimitación correcta de muchos de estos dependerá de la delimitación manual que se realizó sobre los datos originales. De igual manera los elementos que se delimitaron como pertenecientes a alguna clase y que en realidad no eran ENs en la primera fase no podrán ser corregidos por la segunda clasificación, ya que el introducir estos elementos a la fase de entrenamiento para construir el árbol afecta el porcentaje de precisión que se podría alcanzar.

Capítulo 7

Conclusiones y Trabajo Futuro

7.1. Resumen

En muchas ocasiones en la clasificación de entidades nombradas no se aprovecha toda la información que brindan las distintas menciones de una misma EN, aún cuando esta información pudiera servir para apoyar o corregir la clasificación de aquellas menciones que tienen información pobre y que generan clasificaciones erróneas. En este trabajo se presentó un método para aprovechar la información de las distintas menciones de una misma EN mediante el refinamiento de su clasificación inicial. Para esto se experimentó agregando al método de clasificación original dos pasos, uno de vinculación de ENs y otro de refinamiento de la clasificación.

En la primera etapa de la solución, la vinculación de ENs, se propuso un método para unir ENs basadas en la similitud entre ellas, se experimentó con cuatro medidas de similitud, con el objetivo de analizar cuál de éstas era la más adecuada y cual tendría un impacto positivo en el segundo paso.

Para la clasificación de ENs se propuso en primera instancia un conjunto de métodos basados en voto, con el objetivo de integrar las distintas clasificaciones de forma similar a un ensamble y aprovechar la información de estas,

de forma que pudiera incidir en una correcta clasificación final. Además, se analizaron distintos enfoques, buscando de igual forma aprovechar la información de un valor de confianza de las clasificaciones iniciales, tanto para corregir las clases iniciales como para tomarlo como un factor de decisión para la clasificación final. Dentro de estos métodos de voto también se propuso cambiar la distribución de las clases por cadena para que el número y la aparición de clases por cadena se asemejaran de forma más exacta a las cadenas originales y corregir así la aparición de clases incorrectas por cadena.

Con los métodos anteriormente mencionados se alcanzaban mejoras mínimas con respecto a la clasificación inicial, por esta razón se propuso utilizar otro enfoque, un método supervisado, basado en árboles de decisión, pero integrando la información de las menciones de las ENs como atributos para la construcción de los árboles, un método de refinamiento de clasificación utilizando información global, que a diferencia de otros, integra toda la información de las cadenas de ENs además de la que se piensa clasificar.

7.2. Conclusiones

Las fases de delimitación y vinculación de ENs presentaron algunos aspectos que se creé incidieron en el refinamiento de clasificación, uno de ellos fue que aún cuando la medida de similitud elegida fue la de mejor precisión no fue la que presentaba un mejor recuerdo, lo cual reducía el número de ENs que pudieran encadenarse y que pudieran ofrecer mayor información a la clasificación final. También debe considerarse que aún cuando se busca corregir la clasificación inicial de todas las ENs vinculadas, esto no es posible si el primer paso de delimitación fue erróneo, ya que aún cuando se cambie por la clase correcta la delimitación errónea impedirá contarlo como un elemento clasificado correctamente. Sin embargo al entrenar los clasificadores para corregir la clasificación de estos elementos se obtenían resultados menores en términos

de precisión que cuando se clasificaban sin estos elementos, es decir que el introducir estos elementos sólo agregaría ruido en el segundo paso de clasificación.

Los resultados obtenidos bajo los distintos enfoques resultan ser menores o aumentan de forma mínima con respecto a la clasificación inicial, por esta razón se realizó un análisis de los valores ideales, con lo cual se vio que los márgenes de mejora que se tienen son muy pequeños, es decir, los elementos incorrectos que se encontraron vinculados en cadenas de ENs no tienen un impacto importante en los porcentajes de precisión y recuerdo sobre la clasificación inicial, aún cuando se corrigieran todos los elementos incorrectos la mejora no sería sustancial.

Por otra parte los resultados obtenidos varían de acuerdo a los datos que se utilizaron y la fuente de la clasificación inicial, los errores obtenidos en los corpus son distintos, un aspecto a resaltar es que en el corpus desastres los resultados obtenidos son muy buenos, alcanzando una mejora importante con el enfoque de voto, esto debido a las características propias del corpus (como el número de documentos, el número de ENs, el número de ENs por cadena y el dominio más específico de estos), por lo que se creé que teniendo corpus similares los resultados serán parecidos.

Cabe resaltar que el tamaño de los documentos es pequeño en los datos del corpus CoNLL, lo cual en muchos casos limita el número de ENs que se tiene por cadena y por lo tanto aquellas que pueden corregirse, se considera que en el caso de tener documentos más extensos se tendría un mayor número de referencias a una EN y con esto mayor información para su clasificación.

Otro punto a considerar es que las cadenas de ENs que presentaban mayor dificultad al momento de intentar corregirse asignando una sóla clase a todos los elementos de la cadena, fueron aquellas cadenas que en el etiquetado original involucran elementos de tipo lugar y organización, resultando en la clasificación errónea de los elementos que no eran de esa clase; muchas ve-

ces el etiquetado de ciertos elementos puede ser subjetivo y dependerá de la percepción del etiquetador humano, esto sin duda tiene repercusión en la evaluación de la clasificación de las ENs.

Con todo esto se observa que el uso de los métodos presentados en este trabajo y sus resultados dependen de la naturaleza de los datos con los que se trabaja, como pueden ser el tamaño de los documentos, el número de elementos encadenados y el dominio de los documentos, además influyen el método de clasificadoción inicial, los errores de delimitación generados en el primer paso y los errores de la clasificación inicial, estos factores pueden ser limitantes al momento de hacer un segundo proceso de clasificación.

Podemos concluir que en este caso la integración de la información de las vinculaciones de las ENs bajo los distintos enfoques propuestos no representó una mejora notable con respecto a la clasificación inicial, en los casos en los que se mejoraron los valores fue de forma mínima. Aún con esto se concluye que esta primera aproximación de utilizar información global en documentos en español puede servir como base para analizar estos métodos en distintos escenarios, en dominios más específicos y documentos de mayor longitud, donde creemos estos métodos podrían tener una incidencia mayor en el desempeño de la clasificación inicial.

7.3. Trabajo Futuro

Tomando en cuenta los resultados obtenidos por los métodos propuestos en este trabajo, algunas ideas para trabajo futuro se listan a continuación.

- Basados en los resultados de la primera fase de vinculación de entidades creemos que es importante mejorar los resultados en términos de recuerdo y por esta razón se plantea considerar como posibles direcciones el aplicar distintas medidas de similitud y técnicas de agrupamiento para la

vinculación, manteniendo valores con una alta precisión pero que permitan mejorar los valores de recuerdo para conseguir un mayor número de ENs correctas encadenadas.

- En la misma dirección de la primera fase de vinculación se propone utilizar técnicas formales de resolución de coreferencias de ENs para realizar el primer paso de vinculación, se creé que esto podría impactar en la clasificación de las ENs, al tratar de integrar mayor información al proceso de clasificación.

- Utilizar el método de vinculación para tratar de corregir la delimitación de ENs y no enfocarse solamente en la clasificación, al corregir desde el primer paso este proceso podría impactar en el paso de clasificación.

- Con respecto a la segunda fase de refinamiento de clasificación y al observar los porcentajes de mejora alcanzados con árboles de decisión se propone probar con distintos modelos para la clasificación inicial; para analizar la incidencia de otros métodos supervisados en la corrección de la clasificación final.

- En esta misma dirección de la fase de refinamiento de clasificación, se propone experimentar con distintos corpus en distintos idiomas y escenarios, los métodos basados en voto y en árboles de decisión. Por ejemplo, artículos científicos, biología molecular, bioinformática y comunidades médicas, donde se ha generado un importante interés en la clasificación de ENs para identificar nombres de genes.

Índice de figuras

2.1. Ejemplo de árbol de decisión 14

3.1. Taxonomía de trabajos de CEN 24

4.1. Método propuesto 37

5.1. Cadenas verdaderas 49
5.2. Ejemplo cadenas de salida 1 49
5.3. Ejemplo cadenas de salida 2 50
5.4. Ejemplo de vinculación 52

6.1. Ejemplo de voto simple 61
6.2. Ejemplo de voto ponderado 62
6.3. Árbol de decisión con valor confianza MISC 84
6.4. Árbol de decisión con valor confianza ORG 85
6.5. Árbol de decisión con valor confianza PER 86
6.6. Árbol de decisión atributos vecinos: MISC 87
6.7. Árbol de decisión atributos vecinos: ORG 88
6.8. Árbol de decisión atributos vecinos: PER 88
6.9. Voto entre documentos. Método 17 91
6.10. Voto entre documentos. Método 18 91

Índice de Tablas

2.1. Algoritmo ID3 16
2.2. Tabla de contingencia para clase c_i 17

3.1. Resumen de trabajos de CEN 32

4.1. Salida de clasificación inicial 38
4.2. Salida de vinculación 40
4.3. Salida de clasificación final 40

5.1. Algoritmo de vinculación de ENs 44
5.2. Reglas de vinculación 45
5.3. Valores de la métrica B-Cubed de ejemplos 1 y 2 51
5.4. Resultados por medida de similitud 52
5.5. Validación de resultados 57

6.1. Datos de corpus 66
6.2. Resultados de métodos de voto 67
6.3. Análisis de valores de confianza 69
6.4. Análisis de valores de confianza para MISC 70
6.5. Resultados de variantes de voto 75
6.6. Distribución de clases por cadenas 78
6.7. Resultados de enfoques de cadenas 81
6.8. Resultados de árboles de decisión 89

6.9. Resultados de voto entre documentos 92
6.10. Resultados de valores ideales 94

Bibliografía

[1] E. ALAYDIN, *Introduction to Machine Learning*, vol. 1, The MIT Press, 2004.

[2] B. BABYCH AND A. HARTLEY, *Improving machine translation quality with automatic named entity recognition*, in Proceedings of the 11th Conference of the European Chapter of the Association for Computational Lingüistics, Budapest, Hungary, 2003, pp. 1–8.

[3] A. BAGGA AND B. BALDWIN, *Algorithms for scoring coreference chains*, in Proceedings of the Linguistic Coreference Workshop at The Fist International Conference on Language Resources and Evaluation, 1998, pp. 563–566.

[4] D. M. BIKEL, S. MILLER, R. SCHWARTZ, AND R. WEISCHEDEL, *A high performance learning name-finder*, in Proceedings Fifth Conference on Applied Natural Language Processing, March 1998, pp. 194–201.

[5] D. M. BIKEL, R. SCHWARTZ, AND R. WEISCHEDEL, *An algorithm that learns what's in a name*, in Machine Learning, Special Issue on Natural Language Learning, February 1999, pp. 211–231.

[6] W. J. BLACK, F. RINALDI, AND D. MOWATT, *Facile: Description of the ne system used for muc-7*, in Proceedings of the Seventh Message Understanding Conference(MUC-7), 1998.

[7] X. CARRERAS, *Corpus conll2002*, 2002. http://www.cnts.ua.ac.be/conll2002/ner.tgz.

[8] X. CARRERAS, L. MÁRQUEZ, AND L. PADRÓ, *A simple named entity extractor using adaboost*, in Proceedings of the Conference on Computational Natural Language Learning, CONLL2003, May 2003, pp. 152–155.

[9] H. L. CHIEU AND H. T. NG, *Named entity recognition: A maximum entropy approach using global information*, 2002, pp. 190–196.

[10] C. DENICIA-CARRAL, M. M. Y GÓMEZ, L. VILLASEÑOR-PINEDA, AND R. GARCÍA-HERNÁNDEZ, *A text mining approach for definition question answering*, in 5th International Conference on Natural Language Processing (FinTal 2006), August 2006.

[11] J. GE, X. HUANG, AND L. WU, *Approaches to Event-Focused Summarization Based on Named Entities and Query Words*, DUC 2003 Workshop on Text Summarization, (2003).

[12] R. GRISHMAN, *The nyu system for muc-6 or where's the syntax*, in Proceedings of the Sixth Message Understanding Conference, Morgan Kaufmann, 1995, pp. 167–175.

[13] M. HASSEL, *Explotation of named entities in automatic text summarization for swedish*, in Proceedings of 14th Nordic Conference of Computational Lingüistics, Reykjavik, May, 2003, pp. 3–14.

[14] J. HOBBS, J. BEAR, D. ISRAEL, M. KAMEYAMA, A. KEHLER, D. MARTIN, K. MYERS, AND M. TYSON, *Sri international fastus system muc-6 test results and analysis*, in Proceedings of the Sixth Message Understanding Conference(MUC-6), 1995, pp. 237–248.

[15] H. JI AND R. GRISHMAN, *Applying coreference to improve name recognition*, in ACL 2004: Workshop on Reference Resolution and its Applications,

S. Harabagiu and D. Farwell, eds., Barcelona, Spain, July 2004, Association for Computational Linguistics, pp. 32–39.

[16] D. JURAFSKY AND J. H. MARTIN, *Speech and Language Processing*, vol. 1, Prentice Hall, 2000.

[17] Z. KOZAREVA, B. BONEV, AND A. MONTOYO, *Self-training and co-training applied to spanish named entity recognition*, in Proceedings of 4th Mexican International Conference on Artificial Intelligence (MICAI 2005), Lecture Notes in Artificial Intelligence, Monterrey, México, 2005, pp. 770–779.

[18] Z. KOZAREVA, O. FERRÁNDEZ, A. MONTOYO, R. MUÑOZ, A. SUÁREZ, AND J. GÓMEZ, *Combining data-driven systems for improving named entity recognition*, in Data and Knowledge Engineering, Elsevier Science Publishers, 2006, pp. 449–466.

[19] G. R. KRUPKA AND K. HAUSMAN, *Isoquest: Description of the netowl(tm) extractor system as used in muc-7*, in Proceedings of the Seventh Message Understanding Conference(MUC-7), 1998. Published on the website http://www.muc.saic.com/.

[20] A. MANSOURI, L. S. AFFENDEY, AND A. MAMAT, *Named entity recognition approaches*, in International Journal of Computer Science and Network Security, vol. 8, February 2008, pp. 339–344.

[21] T. MITCHELL, *Machine Learning*, McGraw Hill, 1997.

[22] D.ÑADEAU, P. D. TURNEY, AND S. MATWIN, *Unsupervised named entity recognition: Generating gazetteers and resolving ambiguity*, in Advances in Artificial Intelligence, 2006, pp. 266–277.

[23] A.ÑENKOVA AND K. MCKEOWN, *References to named entities: A corpus study*, in Proceedings of the 2003 Conference of the North American

Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of NAACL-HTL 2003 short papers, 2003, pp. 70–72.

[24] L. PADRÓ AND L. MÁRQUEZ, *Herramienta freeling*, 2003. http://garraf.epsevg.upc.es/freeling/.

[25] M. A. PÉREZ-COUTIÑO, *PASCQA: Búsqueda de Respuestas con Base en Anotación Predictiva de Contextos Léxico-Sintácticos*, PhD thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, marzo 2006.

[26] R. SCHAPIRE AND Y. SINGER, *Improved boosting algorithms using confidence-rated predictions*, in Machine Learning, 1999, pp. 297–336.

[27] T. SOLORIO, *Taking Advantage of Existing Named Entity Taggers by Machine Learning*, PhD thesis, National Institute of Astrophysics, Optics and Electronics, september 2005.

[28] T. SOLORIO AND A. LÓPEZ-LÓPEZ, *Learning named entity classifiers using support vector machines*, Lecture Notes in Computer Science: Computational Linguistics and Intelligent Text Processing, CICLing-2004, Springer-Verlag, (2004), pp. 158–166.

[29] T. SOLORIO, A. LÓPEZ-LÓPEZ, L. VILLASEÑOR-PINEDA, AND A. MARTÍNEZ, *Extracting named entities from pseudo speech transcriptions*, in Avances en la Ciencia de la Computación, ENC-2005, 2005, pp. 118–122.

[30] Y. WONG AND H. T. NG, *One class per named entity: Exploiting unlabeled text for named entity recognition*, in Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, India, 2007, pp. 1763–1768.

[31] L. ZHANG, Y. PAN, AND T. ZHANG, *Focused named entity recognition using machine learning*, in Proceedings of SIGIR'2004, 2004, pp. 281–288.

[32] J. ZHU, V. UREN, AND E. MOTTA, *Espotter: Adaptive named entity recognition for web browsing*, in Proceedings of Professional Knowledge Management Conference, Springer-Verlag, Kaiserslautern, Germany, 2005, pp. 518–529.