UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN
INTELIGENCIA ARTIFICIAL

DETECCIÓN Y ORDENAMIENTO TEMPORAL AUTOMÁTICO DE EVENTOS EN TEXTOS DE NOTICIAS

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

PRESENTA:
ALEJANDRO ESTEBAN PIMENTEL ALARCÓN

DIRECTOR DE TESIS:
DR. GERARDO EUGENIO SIERRA MARTÍNEZ
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

CODIRECTORA DE TESIS:
DRA. AZUCENA MONTES RENDÓN
Centro Nacional de Investigación y Desarrollo Tecnológico

CIUDAD UNIVERSITARIA, CD. MX. ENERO 2017

Detección y ordenamiento temporal automático de eventos en textos de noticias

por

Alejandro Esteban Pimentel Alarcón

Ing. Mecatrónico, Universidad Nacional Autónoma de México (2015)

Tesis presentada para obtener el grado de

Maestro en Ciencia e Ingeniería de la Computación

en el

Posgrado en Ciencia e Ingeniería de la Computación

Universidad Nacional Autónoma de México

Ciudad de México. Noviembre, 2016

El esfuerzo de utilizar las máquinas para emular el pensamiento humano
siempre me ha parecido bastante estúpido.
Preferiría usarlas para emular algo mejor.
Edsger Dijkstra

Agradecimientos

Este trabajo fue posible gracias al apoyo del Consejo Nacional de Ciencias y Tecnología (CONACYT) por la beca de manutención 390657 que me permitió realizar los estudios de maestría. (Becario 334720)

Se agradece también el apoyo del CONACYT para los proyectos *Detección y medición automática de similitud textual* con clave 178248 y *Caracterización de huellas textuales para análisis forense* con clave 215179, que se desarrollan en el Grupo de Ingeniería Lingüística (GIL)

Índice general

Índice de figuras

Detección y ordenamiento temporal automático de eventos en textos de noticias

por

Alejandro Esteban Pimentel Alarcón

Resumen

Junto con el aumento de la producción y la velocidad de generación de información textual, crece también el interés por ser capaces de manejarla automáticamente. Una meta ideal del procesamiento automático del lenguaje es lograr, para una computadora, la abstracción artificial de lo que lee, es decir, que una máquina tenga una representación que pueda manejar del significado de lo que se presenta en un documento.

Un paso en dirección hacia la comprensión máquina del lenguaje natural, es lograr el manejo del tiempo. El entendimiento de la secuencia en una narrativa es un aspecto fundamental en la comprensión de un texto.

En esta tesis se muestra el funcionamiento y los resultados de un algoritmo desarrollado para el procesamiento temporal de textos de noticias en español. El enfoque utilizado se basa en dos principales pilares: primero está la detección y el emparejamiento de los eventos con ayuda de etiquetadores sintácticos automáticos, y el segundo está en una doble clasificación de las parejas para encontrar qué relación guardan entre ellas.

La principal aportación de esta tesis para el primer pilar radica en la separación de los pares de eventos según la forma en la que se conectan sintácticamente, con lo que se obtienen tres tipos de pares: se manejan las relaciones neutras o eventos relacionados con el punto de enunciación, las relaciones entre eventos subordinados y las relaciones entre eventos hermanos. Para el segundo pilar, se propone un método de clasificación doble con un entrenamiento semi-supervisado sobre todas las relaciones temporales de Allen [1984] con el apoyo de la representación en vecindades de Freksa [1992].

Los experimentos se separan según las categorías de las parejas de eventos. Para las de eventos subordinados se obtiene una medida-F de 72.5 %. Para las parejas neutras se alcanza una medida-F de 67 %. Mientras que para los verbos hermanos se logra una medida-F de 72 %.

La tesis establece las bases para un corpus etiquetado con información temporal, establece pautas para una representación manejable del tiempo según relaciones temporales en vecindades y se introduce el manejo de una clasificación doble, una inicial gruesa y una segunda especializada para mejorar los resultados de identificación de relaciones temporales.

Detección y ordenamiento temporal automático de eventos en textos de noticias

by

Alejandro Esteban Pimentel Alarcón

Abstract

Along with the increase of production and the rate of generation of textual information, so does the interest to handle it automatically. An ideal goal of automatic language processing is achieving the artificial abstraction of a written input by a computer, that is to say, for a machine to work with a representation of the meaning of what is presented in document.

A step towards machine natural language understanding, is the handling of time. The understanding of the sequence in a narrative is a fundamental aspect in understanding a text.

In this thesis is shown the operation and results of an algorithm developed for temporal processing in news texts in spanish. The approach is based on two main cores: first is the detection and pairing of events with the help of automatic syntactic taggers, and the second is a double classification of coupes to find how they relate to each other.

The main contribution of this thesis for the first core lies in the separation of pairs of events according to the way they are connected syntactically; three types of pairs are obtained: the neutral relations or related to the enunciation point, relations between subordinated events and relations between sibling events. For the second core,it is proposed a method of double classification with a semi-supervised learning approach over the complete set of Allen [1984] temporal relations with the support of the representation in neighborhoods proposed by Freksa [1992].

Experiments were separated according to the categories of the event couples. For the subordinate events an F-measure of 72.5 % was obtained. For neutral couples, an F-measure of 67 % is reached. While for sibling events an F-measure of 72 % is achieved.

The thesis establishes the beginnings of a spanish corpus with temporal information, it also establishes guidelines for a computational representation of time as temporal relationships in neighborhoods and the approach of a two-step classification, a thick initial and a second specialized to improve the identification of temporal relationships.

Capítulo 1

Introducción

La comprensión de la secuencia en una narrativa es un aspecto fundamental en la comprensión de un texto. Las noticias, las novelas, los cuentos, narrativas en general, y muchos otros tipos de textos contienen eventos que ocurren en cierto orden cronológico, ya sea que expresen el tiempo de manera explícita o implícita. El lector necesita obtener de un texto la noción temporal de los sucesos que ocurren en él para poder comprender lo que se está narrando.

Detectar eventos en un texto y ordenarlos cronológicamente de manera automática es un problema de gran interés en varias áreas de investigación. En el área de Procesamiento de Lenguaje Natural (PLN), la detección y el ordenamiento de eventos mejora los procesos de extracción de información, generación de resúmenes y sistemas de preguntas y respuestas, entre otros. En el caso de Lingüística, la detección y ordenamiento de eventos es de gran interés para la comprensión del uso del lenguaje para expresar la percepción del tiempo. En el área de procesamiento semántico, la detección y ordenamiento de eventos puede facilitar la comunicación hombre-máquina de una manera más natural [Allen *et al.*, 2001].

El problema de la detección de eventos ha sido estudiado extensivamente en los últimos años. Actualmente es posible detectar eventos con precisión [Reyes Ortiz, 2013; Saurí *et al.*, 2005; Vargas-Vera y Celjuska, 2004; Reyes Ortiz, 2009]. Sin embargo, en el procesamiento de la información temporal, ordenar dichos eventos únicamente se ha logrado en escenarios restringidos como oraciones simples, telegramas o itinerarios [Saquete *et al.*, 2006; Bramsen *et al.*, 2006a; Allen y Koomen, 1983].

Mas recientemente varios enfoques han intentado ordenar eventos en textos más generales

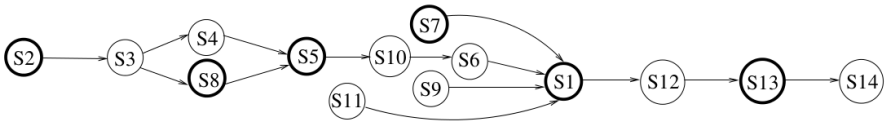
[Mani *et al.*, 2003; Filatova y Hovy, 2001]. Los que tienen un enfoque más cercanos a esta tesis son los de Chambers *et al.* [2014], Bramsen *et al.* [2006b] y Kolomiyets *et al.* [2012]. Estos trabajos se orientan siempre para el inglés y comparten una reducción de complejidad en el ordenamiento, lo cual causa una pérdida de información, y en consecuencia depender de una detección previa de *eventos relevantes* para el ordenamiento. En la figura 1-1 se muestran dos ejemplos que muestran la forma en la que se ha manejado el ordenamiento en el estado del arte.

Las investigaciones que abordan el análisis del tiempo en el área de Procesamiento del Lenguaje Natural, están enfocadas a encontrar dentro del texto palabras, frases o indicadores que establezcan una ubicación temporal de las expresiones. Es decir, se busca encontrar qué elementos son los que se usan en el lenguaje para expresar momentos o sucesiones, y que las computadoras los puedan utilizar de forma estandarizada. Una de las formas más claras en la que se puede representar un punto o un intervalo de tiempo es mediante fechas, es decir, expresiones numéricas que representan días, horas, minutos, segundos, etc. Estas expresiones son localizadas y traducidas en una notación que las computadoras pueden entender, esto es, en etiquetas temporales [Saquete *et al.*, 2003; Mani y Wilson, 2000; Wilson *et al.*, 2001; Saquete *et al.*, 2006; Filatova y Hovy, 2001].

Se ha realizado también un gran número de investigaciones, principalmente teóricas, en el uso de relaciones temporales para establecer una cadena de puntos de referencia, comprendidos principalmente por eventos anclados a una punto temporal conocido o a otro evento, con ayuda de las cuales se pueden armar líneas de tiempo. Este tipo de enfoque permite una representación más abstracta del tiempo que se asemeja más a la forma en la que los humanos se comunican, puesto que se apoyan más en expresiones relativas que absolutas, usando muchas veces como referencia el momento del habla o de creación del documento. Las relaciones temporales pueden ser usadas de manera semejante [Freksa, 1992; Allen y Koomen, 1983; Allen, 1983; Bramsen *et al.*, 2006b; Chambers *et al.*, 2014; Vilain y Kautz, 1986].

1.1. Problemática

A partir de un texto en español se pretende detectar los eventos que contenga y obtener el orden cronológico que hay entre dichos eventos. No obstante:

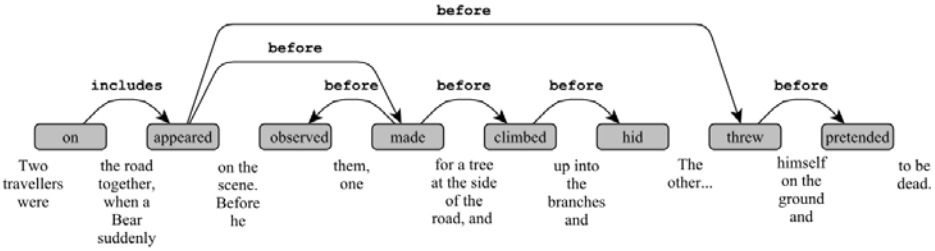


S1	A 32-year-old woman was admitted to the hospital because of left subcostal pain...
S2	The patient had been well until four years earlier,
S5	Three months before admission an evaluation elsewhere included an ultrasonographic examination, a computed tomographic (CT) scan of the abdomen...
S7	She had a history of eczema and of asthma...
S8	She had lost 18 kg in weight during the preceding 18 months.
S13	On examination the patient was slim and appeared well. An abdominal examination revealed a soft systolic bruit... and a neurologic examination was normal...

(a) Ejemplo de identificación de eventos y ordenamiento por grafos [Bramsen *et al.*, 2006b].

En la imagen se muestra un grafo con una serie de nodos que representan la información de interés a ser ordenada. La naturaleza del grafo permite que no haya incongruencias en las relaciones, pero también las limita únicamente a relaciones de *.ªntes de"* y *"despues de"*.

Imagen tomada de *Inducing temporal graphs* [Bramsen *et al.*, 2006b].



(b) Ejemplo de identificación de relaciones temporales dentro de una oración [Kolomiyets *et al.*, 2012].

En la figura se muestra una oración en la que se hace una identificación de acciones y se encuentran relaciones temporales entre ellas. Este ordenamiento permite una mayor cantidad de relaciones.

Imagen tomada de *Extracting narrative timelines as temporal dependency structures* [Kolomiyets *et al.*, 2012].

Figura 1-1: Ejemplos de ordenamiento temporal en el estado del arte.

En la figura se muestran ejemplos de los trabajos de Bramsen *et al.* [2006b] y Kolomiyets *et al.* [2012] para el ordenamiento temporal de eventos.

- Los textos contienen información no estructurada, es decir, información que no está organizada según el tipo de datos que contiene, no tiene el orden que se encontraría en una base de datos para que una máquina la pueda interpretar.
- Los reconocedores de eventos son sistemas especializados, enfocados en objetivos particulares, incompatibles con el idioma español.
- Los estudios lingüísticos del aspecto gramatical, el cuál es una propiedad de los verbos que marca la percepción de duración de los mismos y su estructura temporal, no han formalizado un modelo computacional.
- No se cuenta previamente con los documentos etiquetados adecuadamente para entrenar a una computadora en la obtención de los resultados que se buscan.

1.2. Objetivos

El objetivo principal es lograr que a partir de textos español se detecten automáticamente los eventos que contiene el documento y se obtenga el orden cronológico que hay entre dichos eventos.

El sistema a desarrollar entonces debe ser capaz de recibir como entrada un texto, y obtener como salida una representación de la información temporal que en el texto se presenta, tomando en cuenta cada uno de los hechos o eventos que se mencionan.

Objetivos específicos

- Determinar una representación del tiempo que tome en cuenta eventos o sucesos narrados.
- Detectar automáticamente en textos de noticias cada uno de los eventos que se mencionen cuya duración pueda tener una representación temporal.
- Dar una estructura ordenada a la duración los eventos encontrados según la representación temporal hallada sin pérdidas por simplificaciones.
- Desarrollar un programa de ordenamiento temporal que sea capaz de trabajar con textos en español.

1.3. Hipótesis

Dentro del ámbito de el procesamiento temporal como relaciones entre eventos existen diferentes propuestas para trabajar con las relaciones temporales. Una que se ha encontrado particularmente interesante es la estructura propuesta por Christian Freksa [1992].

Se parte de la idea de que se puede implementar dicha estructura dentro un desarrollo computacional para lograr un mejor manejo de la información temporal por parte de una computadora.

1.4. Motivación

Procesar información de manera automática es de gran interés en una amplia gama de áreas de investigación donde se busca una comunicación natural entre humanos y maquinas. Aplicaciones como *Siri* y *Cortana* son ejemplos de la popularidad de estas áreas. Otros ejemplos incluyen sistemas de preguntas y respuestas, organización y vinculación de documentos y creación de resúmenes. Todas estas áreas pueden mejorar su rendimiento con la capacidad de detectar y ordenar eventos de manera automática.

El manejo de información temporal y su ordenamiento automatizado puede ayudar a las tecnicas de extracción de información temporal. De igual manera ayudaría a sistemas de recuperación de información y pregunta-resupuesta.

1.5. Contribuciones principales

La principal contribución de esta tesis es el método que permite dar un ordenamiento cronológico de manera automática a los eventos de textos en español.

Otras de las principales contribuciones de esta tesis se encuentran en el enfoque de relaciones temporales, puesto que:

- Se propone un emparejamiento automático de los eventos entre los que se buscará una relación temporal.
- Se desarrolla un sistema que trabaja con todas las relaciones temporales que se pueden presentar entre dos eventos, según lo expone Allen [1983].

- Se utiliza en el sistema creado una clasificación basada en representaciones versátiles propuestas por Freksa [1992], que permiten la representación de vaguedad y conocimiento incompleto sobre las relaciones.

1.6. Estructura de la tesis

En el capítulo 2 se hace mención a los trabajos relacionados con el desarrollo de esta tesis, se separan en dos temas principales, la identificación y manejo de los eventos de un texto por un lado, y por otro lado se citan los trabajos relacionados con el análisis temporal, en dicho análisis temporal se hace principal hincapié, pues es hacia donde está más enfocada esta tesis. En el capítulo 3 se enlistan todos los conceptos y antecedentes teóricos necesarios para la correcta comprensión del resto del documento. En el capítulo 4 se desarrolla a profundidad el método propuesto para el procesamiento del tiempo, desde la etapa de pre-procesamiento de los documentos, la extracción de la información de los eventos, así como la separación que se hace para su manejo independiente, y por último las consideraciones que se tuvieron para la clasificación de las relaciones. En el capítulo 5 se muestran todos los pasos del capítulo 4 aplicados al caso de los textos de noticias en español, y se muestran también los resultados obtenidos acompañados con una evaluación. En el capítulo 6 se muestran las conclusiones que se pueden obtener de los resultados de los experimentos, así como trabajo a futuro y potencial para el sistema.

Capítulo 2

Estado del arte

El principal interés de este trabajo de tesis, es el manejo y ordenamiento temporal de los hechos que narra un texto. Por lo tanto, es fundamental tener o hallar los eventos que tendrán una presencia temporal.

En este capítulo se mencionan diferentes investigaciones que han trabajado con alguno de los temas de detección, extracción y ordenamiento de eventos a partir de textos de manera automática. Más adelante, se arma una tabla comparativa entre las características de los sistemas abarcados.

2.1. Eventos

Los eventos han tenido muchas definiciones según el área que los estudie y el enfoque con el que se manejen. Los textos se pueden ver como sucesos que involucran progresión, como cambios de lugar, de estado, de foco temporal u otros tipos de modificaciones en diferentes patrones. [Bramsen *et al.*, 2006b; Allen y Ferguson, 1994].

Sin embargo, los eventos también se han representado como instantes o intervalos que tienen consecuencias y dan lugar a otros eventos. Entonces, se puede definir un evento según su causalidad y espacio-tiempo [Galton y Augusto, 2002; Sowa, 1999].

Otros enfoques toman al evento como el núcleo de una estructura que tiene características de espacialidad, temporalidad, causalidad, protagonismo e intencionalidad [Reyes Ortiz, 2013; Zwaan *et al.*, 1995].

Cuando se trata de aplicaciones, el manejo de los eventos suele hacerse de forma más práctica. Dentro de *TimeML*, un conjunto de reglas que utiliza para codificar documentos electrónicamente, Pustejovsky *et al.* [2002] definen al evento como aquellas expresiones que participan en la narrativa que pueden ser ordenadas temporalmente. Saurí *et al.* [2005] utilizan la definición anterior para desarrollar su identificador de eventos *Evita*, que se enfoca en verbos, sustantivos, frases nominales y frases adjetivas.

Otros trabajos más enfocados al procesamiento de eventos utilizan herramientas como CLAUSE-IT o CONTEX [Hermjakob y Mooney, 1997] para identificar estructuras sintácticas que puedan ser usadas como eventos; para ello, emplean las cláusulas y los verbos que conforman las oraciones compuestas y los utilizan como representación de los eventos [Mani *et al.*, 2003; Filatova y Hovy, 2001].

Dentro de esta tesis, se maneja el enfoque que considera al evento como núcleo de una estructura sintáctica [Reyes Ortiz, 2013; Zwaan *et al.*, 1995]. Un evento se puede ver como un verbo, y la característica en la que nos enfocaremos será en la de temporalidad, ocurren dentro de un intervalo de tiempo, sobre el cual se puede distinguir un inicio y un fin.

2.2. Análisis temporal

Los eventos son fundamentales para el estudio del tiempo. El análisis temporal es el estudio de la información semántica de un texto que se refiere al tiempo. Dentro de esta rama existe una evaluación particularmente importante, el *TempEval* [Verhagen *et al.*, 2010, 2007] que propone un modo simple para evaluar la extracción de información temporal.

En su primera versión, el *TempEval* consistió en tres tareas para identificar la relación temporal entre una lista de eventos previamente localizados[Verhagen *et al.*, 2007].

1. La interpretación de la relación temporal entre dos eventos dados dentro de una misma oración.

2. La interpretación de la relación temporal entre un evento dado y el tiempo de creación del documento.

3. La interpretación de la relación entre dos eventos dados en oraciones adyacentes.

Más adelante, para la segunda versión del *TempEval*, además de mantener las mismas tareas que en la primera versión se agregaron las tareas de identificación de eventos y la identificación de expresiones temporales, es decir, palabras o expresiones que denoten un punto particular en el tiempo como fechas, días y horas en concreto [Verhagen *et al.*, 2010].

El *TempEval* busca obtener más información de manera automática. Para esto, la diferencia entre las primeras dos versiones radica en la opción de efectuar las tareas sobre un texto de entrada sin ninguna información preprocesada, a diferencia de las versiones anteriores en las que cuentan con etiquetas humanas para los eventos, además de que se utilizan más relaciones temporales [UzZaman *et al.*, 2013].

Una de las diferentes formas utilizadas para expresar el tiempo en el que ocurre un evento es mediante la especificación explícita de la fecha en que dicho evento ocurre. Han surgido varias maneras para anotar el tiempo en un intento de normalizar la información temporal que se puede extraer a partir de textos. Los diferentes esquemas de anotación definen reglas guía para que una referencia temporal tenga siempre la misma representación.

Uno de los primeros esquemas de anotación temporal de gran importancia es el esquema *TIDES* [Ferro *et al.*, 2001], surge para asignar etiquetas con fechas para las expresiones donde un humano puede determinar el valor del tiempo al que se refiere la expresión. Complementando esto, el esquema *STAG* [Setzer, 2001] propone una ampliación en el sistema de etiquetado y ser capaz de tomar en cuenta expresiones que relacionaran eventos entre sí y no solamente eventos con fechas. Por su parte *TimeML* [Pustejovsky *et al.*, 2002] establece líneas guía para anotar eventos, expresiones temporales y relaciones para ambos casos, además de agregar nuevas características como la capacidad de dar a un evento múltiples relaciones. El lenguaje de etiquetas *TimeML* fue el adoptado por el *TempEval* desde el 2010 [Group *et al.*, 2009].

Saquete *et al.* [2003] trabajan un sistema taxonómico de expresiones temporales para su sistema *TERSEO* que clasifica las expresiones encontradas según si son explícitas o implícitas. Por un lado, se consideran explícitas si se encuentran en el texto expresiones de fechas completas, como *"11/01/2002"* o *"4 de enero del 2002"*. Por el otro, se consideran implícitas si se pueden encontrar por hacer referencia a otra fecha conocida, tal es el caso de palabras clave como *"esta Navidad"* o *"el próximo mes"*. De manera independiente, dicho sistema también clasifica con base en el tipo de expresión temporal, según su grado de vaguedad; por ejemplo una fecha

concreta que se define perfectamente con expresiones como *"ayer"*, que puede ser representada por un día con un formato *dd/mm/yyyy*. También puede ser una expresión que se refiera a un intervalo, con un formato *[dd/mm/yyyy - dd/mm/yyyy]*, que se da con expresiones como *"la semana pasada"*; o bien, expresión difusa, que da un intervalo aproximado de tiempo con expresiones como *"hace unos días"*.

Mani y Wilson [2000] utilizan para su etiquetado una notación estándar ISO del calendario para denotar siglo, año, mes, día, hora, minuto y segundo (*CC:YY:MM:DD:HH:XX:SS*) de las expresiones que manejan. Su enfoque está dirigido a frases dentro del texto que denotan tiempo, más que eventos en sí, palabras como nombres de meses o referencia a días o temporadas del año.

Por su parte, Filatova y Hovy [2001] utilizan su propia representación del tiempo en etiquetas que les permiten expresar intervalos entre dos fechas o puntos límite de los tiempos en los que pudieron ocurrir los eventos.

Jung y Stent [2013] se enfocan en el uso de características léxicas para hacer la anotación temporal y desarrollan *ATT1*. Defienden que una representación léxica más rica, utilizando ventanas de contexto más grandes, pueden suministrar la información necesaria para hacer el análisis temporal sin tener que recurrir a capas más profundas de representación de información de un texto, como los análisis sintácticos profundos o el manejo de características semánticas.

Hagège y Tannier [2007] presentan el sistema *XRCE-T* para el tratamiento de expresiones temporales. Su trabajo plantea tres niveles de procesamiento. El primer nivel se presenta de manera local, toman en cuenta ventanas contextuales y reglas que les dan información según las expresiones que se utilicen, como fechas y otras marcas temporales. El segundo nivel consiste en el nivel de oración, para ello aplican una búsqueda para detectar los vínculos entre expresiones temporales y eventos a los que pueden estar afectando. El último nivel de procesamiento se hace con respecto al documento, en éste se comparan todas las expresiones que estén referidas con respecto al tiempo de creación del documento.

Strötgen y Gertz [2010] desarrollan el conjunto de reglas para la extracción y normalización de expresiones temporales *HeidelTime*. Principalmente utilizan expresiones regulares y analizadores lingüísticos para la normalización. Este sistema fue el que obtuvo el mejor desempeño dentro del *TempEval* para la extracción de expresiones temporales.

Las relaciones temporales también se consideran dentro de las aplicaciones y los concursos del área, como el *TempEval*. Dentro del lenguaje *TimeMl* también se incluyen formas para anotar relaciones que se presenten entre los eventos [Pustejovsky *et al.*, 2002]. Sin embargo, la forma en la que se pueden relacionar eventos en el *TempEval* se limita a *BEFORE*, *AFTER*, *OVERLAP* y *VAGUE* [Verhagen *et al.*, 2010].

Bramsen *et al.* [2006a] utilizan grafos direccionados para analizar plantillas médicas. Localizan fragmentos de texto que no presentan saltos de tiempo [Webber, 1988]. El método consiste en ordenar segmentos temporales que definen como un fragmento de texto que no exhibe un cambio abrupto de foco temporal, y utilizan aristas dirigidas para representar el orden cronológico. Las aristas dirigidas del grafo ordenado son una restricción topológica que considera como un punto fuerte en el ordenamiento temporal; dan como ejemplo la prohibición de ciclos en el grafo, pues el tiempo corre en una sola dirección.

UzZaman y Allen [2010] desarrollan los sistemas *TRIPS* y *TRIOS*, ambos sistemas utilizan una combinación de análisis semánticos profundos, redes lógicas de Markov y clasificadores de campos aleatorios. Estos sistemas participaron desde el *TempEval 2* y fueron de los primeros sistemas en manejar como entrada un texto crudo sin etiquetas previas.

Chambers *et al.* [2014] utilizan pares de eventos y los procesan por módulos especializados que intentan relacionar todos los pares posibles de eventos dentro de un texto. La intención es que todos los eventos del documento queden conectados con alguna relación. Además, hacen énfasis sobre la importancia de que el sistema encuentre los eventos a relacionar, de modo que no quede restringido, como en el *TempEval*, por los grupos de eventos que los anotadores consideren que tienen relaciones relevantes. El texto es manipulado de forma serial por clasificadores de relaciones, cada pasada toma la salida de la anterior, los módulos están ordenados de manera que el más preciso es el primero y posteriormente se vayan corrigiendo eventos que presenten casos particulares de error. Para el trabajo, los autores utilizan un corpus especializado llamado *TimeBank-Dense* [Cassidy *et al.*, 2014] el cual tiene notas para relacionar distintos casos:

- Evento-Evento, Evento-Tiempo, y Tiempo-Tiempo en la misma oración (tiempo se refiere a una marca temporal encontrada en el texto, como una fecha).

- Evento-Evento, Evento-Tiempo y Tiempo-Tiempo entre una oración y la siguiente.

- Evento-DCT para todo el documento (donde DCT es el tiempo de creación del documento por sus siglas en inglés: *document creation time*).
- Tiempo-DCT para todo el documento.

El etiquetado incluye las etiquetas: *BEFORE*, *AFTER*, *INCLUDES*, *IS INCLUDED*, *SIMULTANEOUS* Y *VAGUE*. De las cuales, *VAGUE* ocupa el 46 % del grafo anotado, esto es una medida empírica de cuántos nodos no tienen una semántica clara.

La disminución de relaciones que se toman en cuenta en los procesos automáticos, en comparación con las relaciones de Allen [1983] ya mencionadas, representa gran aumento de simplicidad de etiquetado y procesamiento; sin embargo, también representa pérdida de información.

Kolomiyets *et al.* [2012] analizan textos más generales: cuentos para niños. Utilizan dos modelos de análisis, y comparan las salidas que obtienen los analizadores. Su trabajo propone una extracción de información que identifique una única linea de tiempo para un texto. Al igual que los otros trabajos mencionados, se limitan a formas simples en las que pueden estar ordenados los eventos. Se basan en la definición y notación de eventos del *TimeML*, pero se separa del concurso *TempEval* por considerar los eventos que relaciona muy limitados. Proponen también una nueva estructura de relación entre eventos que se basa en una dependencia temporal que se une mediante las relaciones: *BEFORE*, *AFTER*, *OVERLAP* o *IDENTITY*. Conservan una mayor cantidad de información al utilizar una nueva forma de caracterizar estructuras de tiempo con árboles de dependencias temporales.

Llorens *et al.* [2010] presentan el proyecto TIPSem, un sistema para extraer información temporal en ingles y en español. Para sus modelos de aprendizaje automático, utilizan datos de diferentes niveles de análisis lingüístico, de los cuáles se enfocan principalmente en características semánticas. Este trabajo obtuvo los mejores resultados en el *TempEval 2* y fue utilizado como un sistema base para la siguiente versión del *TempEval*.

Chambers [2013] desarrolla en sistema *Navytime* para manejar textos crudos en inglés dentro de la competencia del *TempEval*. Participa en la extracción de eventos con una clasificación binaria para cada una de las palabras de un texto; adopta el enfoque basado en reglas para la extracción de expresiones temporales; y por último, utiliza una combinación de reglas y aprendizaje máquina para establecer las instancias que deben ir relacionadas según la notación de los corpus de entrenamiento.

Filannino *et al.* [2013] desarrollan el *ManTIME*, un trabajo para el *TempEval* en el que investigan el cambio de desempeño con respecto a diferentes tipos de características a considerar. En particular, muestran una influencia negativa al utilizar características basadas en WordNet. También muestran un entrenamiento combinado utilizando dos corpus de entrenamiento proporcionados por *TempEval*, el primero de los corpus anotado por humanos y el segundo anotado con apoyo de un sistema automático; el segundo corpus no parece tener ninguna influencia positiva al ser usado como entrenamiento.

Bethard [2013] desarrolla el sistema ClearTK-TimeML para el *TempEval*. Se trata de un sistema minimalista que trabaja con un conjunto pequeño de características morfo-sintácticas. La estrategia consiste en utilizar una cadena de modelos de aprendizaje automático para cada una de las tareas en ingles del *TempEval*. Este sistema fue el ganador de la tarea de relaciones temporales.

Yoshikawa *et al.* [2009] utilizan la misma estructura de tareas propuesta en el *TempEval* para obtener relaciones temporales entre eventos, expresiones temporales, y un evento y una expresión temporal. Sin embargo, proponen un enfoque diferente para identificar el tipo de relación, en el cual, en lugar de separar las relaciones según estas clases, utilizan un modelo de lógica de Markov para identificar el tipo de relación de todas las categorías a la vez. Según reportan en sus resultados, esta forma de unir las relaciones aumenta en 2 % la precisión del tipo de relaciones.

En la figura 2-1 se presenta una tabla comparativa de algunos de los sistemas antes mencionados. Se muestran los resultados reportados para la extracción de eventos, la extracción de expresiones temporales y la categorización de las relaciones temporales. En la última columna se mencionan las características principales del enfoque para cada sistema.

Sistema	Eventos	Expresiones	Relaciones	Enfoque
TERSEO	–	88.92 %	–	Clasificación según referencia temporal en explícitas e implícitas. Clasificación según representación en concretas, puntos y difusas.
ATT1	81.16 %	80.71 %	–	Aumento de contexto en lugar de complejidad de información.
XRCE-T	79.00 %	69.00 %	51.00 %	Análisis sintáctico profundo. Tres niveles de análisis: local, oración y documento.
HeidelTime	86.00 %	85.00 %	–	Sistema basado en reglas. Expresiones regulares para las expresiones temporales.
TRIPS	67.00 %	85.00 %	57.50 %	Análisis semántico profundo. Redes lógicas de Markov.
TRIOS	77.00 %	85.00 %	58.50 %	Clasificadores de campos aleatorios.
TIPSem	91.00 %	83.00 %	59.00 %	Entrenamiento automático sobre información semántica.
NavytIme	80.30 %	72.43 %	46.80 %	Sistema híbrido de reglas y aprendizaje automático. Clasificación dividida en clasificadores especializados.
ManTIME	85.00 %	84.00 %	–	Campos aleatorios condicionales. Investigación de desempeño con respecto a diferentes campos de características.
ClearTK	77.30 %	82.70 %	31.00 %	Enfoque minimalista. Esquema de características morfo-sintácticas.

Figura 2-1: Tabla comparativa de resultados

En la figura se muestra una tabla con la comparación de los resultados y enfoques entre sistemas de análisis temporal.

Capítulo 3

Marco teórico

El tema desarrollado en esta tesis se encuentra dentro del área de la lingüística aplicada y, más particularmente, dentro de la lingüística computacional. Herramientas computacionales trabajan en conjunto con la teoría lingüística para dar lugar al desarrollo de tecnologías en las que se involucra el lenguaje.

Dentro de este capítulo se presentan las bases teóricas de los análisis lingüísticos utilizados para el desarrollo de esta tesis. De igual manera, se exponen las técnicas y métricas computacionales que se utilizan como herramienta para implementar los algoritmos. Por último, se explica la teoría especializada para el manejo particular de información temporal.

3.1. Análisis lingüístico

La lingüística es la ciencia que estudia el lenguaje natural. Estudia la estructura general de diversos lenguajes naturales con el fin de descubrir leyes universales que rijan su comportamiento [Bolshakov y Gelbukh, 2004]. Es importarte tener presentes los conceptos de las ramas principales que la conforman. Dichos conceptos, según los definen Bolshakov y Gelbukh [2004], se presentan a continuación:

Fonología: Se encarga de los sonidos que comprenden el discurso, tomando en cuenta todas las similitudes y diferencias que permiten formar y distinguir palabras.

Morfología: Se encarga de la estructura interna de las palabras, así como de las leyes que con-

ciernen a la formación de nuevas palabras a partir de partes menores que las constituyen.

Sintaxis: Se encarga de considerar las estructuras de las oraciones y las maneras en las que las palabras se conectan entre sí para formarlas.

Semántica y pragmática: Dos niveles de estudio que se encuentran muy relacionados. La semántica se encarga de las palabras y los textos, mientras que la pragmática estudia las motivaciones de las personas para producir oraciones específicas en situaciones específicas.

Para el procesamiento de textos de forma automática, la lingüística computacional ha desarrollado herramientas que tienen sus bases en las técnicas y los estudios lingüísticos. Existen diferentes programas de computadoras que agrupan algoritmos desarrollados con el fin de hacer procesos lingüísticos comunes, entre los cuáles se encuentran:

Stanford Natural Language Processing Group: Cuenta con diferentes módulos con herramientas especializadas. Funciona con el lenguaje de programación *Java* y cuenta con la posibilidad de analizar muchos idiomas, aunque está enfocado principalmente en el inglés.

Freeling: Es un paquete de análisis del lenguaje, liberado bajo la licencia GNU de la fundación de software libre. Esta diseñado para funcionar como una librería para lenguajes de programación como C++ y *Python*. Cuenta con soporte para la mayoría de las lenguas europeas y trabaja con las etiquetas EAGLES para representar partes de la oración.

TnT: Un etiquetador de partes de la oración enfocado principalmente para el inglés y el alemán, aunque también tiene soporte para otros idiomas, incluyendo el español.

TreeTagger: Es una herramienta para la notación de partes de la oración. Trabaja con idiomas tan variados como Alemán, Inglés, Francés, Italiano, Español, Ruso, Portugués, Chino, Latín, entre otros.

Cada herramienta trabaja con la información lingüística de diferente forma. Sin embargo, comparten los objetivos fundamentales que tienen los estudios lingüísticos para obtener información. Los principales procesos a los que se someten los textos se describen a continuación.

3.1.1. Segmentación

El primer paso de los procesamientos automáticos de análisis lingüístico es la segmentación o *tokenización*. Esto se refiere a la tarea de diferenciar y separar las las palabras que se encuentren dentro de un texto [Jurafsky y Martin, 2007]. La idea más simple sería buscar espacios y obtener todo lo que hay entre ellos como palabras; sin embargo, este enfoque no funciona para situaciones en las que se tienen signos de puntuación u otros símbolos entre espacios. Además, los segmentos se diferencian de las palabras al poder estar formados por más de una palabra, por ejemplo, "Estados Unidos" se puede considerar como un solo segmento.

Los estudios han logrado obtener listas de reglas capaces de trabajar con un gran número de casos y excepciones. De esta forma, se pueden reconocer automáticamente no solo palabras, sino cifras numéricas, fechas, términos comunes de varias palabras, entre otras cosas.

3.1.2. Análisis morfológico

La morfología es el estudio de la forma en la que las palabras están construidas a partir de unidades más pequeñas que mantienen un significado, dichas partes son llamadas morfemas [Jurafsky y Martin, 2007].

Para muchos procesos del PLN, es necesario obtener información acerca de palabras que tengan una misma raíz en común, un ejemplo muy claro de esto nos lo dan los buscadores. Dentro de un texto, los verbos aparecen conjugados de formas diferentes, pero si se quiere buscar todas las apariciones de un verbo en particular sin importar la forma en la que está expresado, es necesario convertir las palabras tal cual se encuentran en el texto, en una forma estandarizada que las unifique.

El proceso de convertir una palabra a la forma estándar con la que se representa (para los verbos, esto comúnmente es su forma infinitiva) de manera automática, es el proceso conocido como *lematización*. La forma raíz o estándar con la que se representa la palabra, es conocida como *lema* [Jurafsky y Martin, 2007].

Etiquetado *PoS*

Una oración es un elemento básico del lenguaje; expresa un pensamiento completo y está formado por palabras, que ya sea individualmente o en combinación con otras palabras, re-

presentan ideas. Diferentes ideas, objetos, acciones, relaciones entre ideas, requieren diferentes representaciones. Estas diferentes representaciones son llamadas, de forma colectiva, partes de la oración [Harris, 1985].

Las partes de la oración entre las que se pueden clasificar las palabras dependen de la gramática que se utilice para el análisis. Sustantivos, verbos, adjetivos y otras categorías importantes, están presentes en todas las gramáticas; sin embargo, cada una consta de sub-categorías o accidentes de acuerdo con la clase de palabra que se trate, éstas pueden ser conjugaciones, género, número, entre otros.

El análisis morfológico consiste en detectar a qué categoría o categorías pertenece una palabra determinada dentro de un texto, según las reglas gramaticales que se usen.

El etiquetado de partes de la oración (*PoS* por sus siglas en inglés) es la parte del análisis en el que se asigna una categoría a cada palabra o segmento del texto dependiendo de la función que éste desempeñe en una oración. Este etiquetado es muy importante para lograr obtener características sofisticadas de un texto.

En el ejemplo 3.1.1 se muestra un caso en el que una misma palabra puede caer en dos categorías básicas diferentes, como sustantivo o como verbo. La forma de diferenciar entre ambas se logra no sólo por su forma, sino por la parte de la oración en la que se encuentra.

Ejemplo 3.1.1 Una misma palabra puede tener diferentes categorías, la *adecuada* se conocerá dependiendo del contexto, pero las opciones que tiene para tomar son fijas.

traje : Del verbo *traer* en primera persona singular (yo) pretérito indicativo.

traje : Vestido completo de una persona. Conjunto masculino de chaqueta, pantalón y a veces chaleco

3.1.3. Análisis sintácticos

Para que una oración pueda expresar un pensamiento completo, necesita tener un sujeto y un predicado. El sujeto es de lo que está hablando una oración, mientras que el predicado es lo que se dice sobre el sujeto dentro de la oración [Harris, 1985].

El análisis sintáctico tiene como principal función determinar la estructura de los textos de entrada. En particular, un análisis sintáctico debe ser capaz de identificar el sujeto y los objetos

de cada verbo y determinar el componente afectado por cada palabra o frase modificadora [Grishman, 1986].

Un ejemplo sencillo de esto se aprecia en el ejemplo 3.1.2, una oración se muestra dividida por sus partes sintácticas simples; las palabras se agrupan en módulos sintácticos que aportan cierta información.

Ejemplo 3.1.2 Considere la oración: "La empresa recortó sus gastos en agosto."

Un análisis básico nos puede mostrar sus principales componentes sintácticos en la figura 3-1.

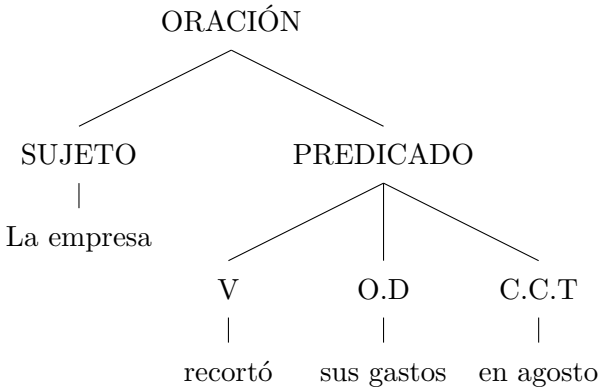


Figura 3-1: Análisis sintáctico

La oración se divide en sus componentes básicos, sujeto y predicado. El sujeto contiene el elemento del cuál se está hablando, mientras que el predicado contiene lo que se está diciendo del elemento. A su vez, el predicado tiene diferentes partes como el objeto directo (O.D), aquello en lo que recae la acción, y el complemento circunstancial de tiempo (C.C.T), es decir, el momento en el que ocurre la acción. Este enfoque es muy simplista, las reglas de las gramáticas varían en la fineza y en la complejidad que pueden tomar las divisiones.

Dependencia sintáctica

Un grupo de palabras que juntas contienen un sujeto y un predicado se le llaman cláusulas u oraciones simples [Harris, 1985]. Como consecuencia, cada cláusula necesita uno y solo puede tener un verbo[1]. Cuando una oración completa tiene un solo verbo, se trata de una oración simple. Cuando se tiene más de un verbo en una oración, se trata de una oración compuesta.

[1]Los núcleos de las oraciones muchas veces se presentan como frases o perífrasis verbales, que pueden contener más de un verbo, sin embargo se refieren a una sola acción. Por ejemplo: *romper a reir*, *echarse a llorar*. *ir caminando*. Sin embargo, siguen siendo un solo núcleo, y solo en uno de estos verbos recae la carga semántica: *reir*, *llorar*, *caminando*

Frecuentemente, las partes sintácticas de una oración están formadas por oraciones simples o cláusulas, esto da lugar a que la oración principal tenga más de un verbo. En estos casos, se dice que la cláusula o la oración simple está subordinada a la oración principal. En el ejemplo 3.1.3 se puede ver un ejemplo sencillo de cómo una oración forma parte de otra formando una oración compuesta.

Ejemplo 3.1.3 Considere la oración: "La empresa no hizo buenas inversiones mientras el dolar estaba subiendo ."

La división de la oración en las partes que la componen se expone en la figura 3-2.

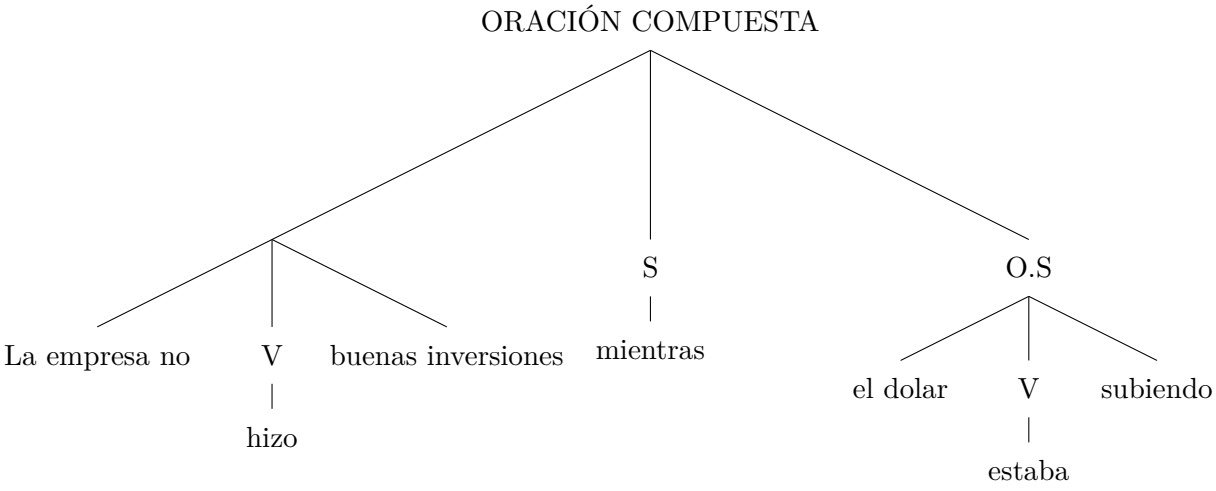


Figura 3-2: Oración compuesta

La oración compuesta cuenta con dos verbos (V), uno de los cuáles es el verbo principal de la oración y el segundo es el núcleo de una oración subordinada (O.S). La unión está dada por un elemento subordinante (S) que hace que la subordinación tenga la función de un complemento circunstancial.

En general, los textos son estudiados utilizando una representación de las partes que conforman a una oración, como sujeto y predicado, y todas las sub-categorías sintácticas según sea el análisis hasta llegar a los núcleos de cada parte de la oración.

No obstante, también se puede hacer un análisis de dependencias sintácticas, es decir, encontrar las palabras que cumplen jerárquicamente funciones dentro de la oración. En el caso de las gramáticas de dependencias, el verbo suele ocupar la posición más alta de la estructura, dando lugar a componentes como sus *sujetos*, *objetos*, *complementos*, entre otras categorías y sub-categorías dadas por la gramática.

En la figura 3-3 se muestra el ejemplo de un análisis sintáctico automático de la oración: *Vinken se unirá al consejo como director*. En la figura 3-3a se muestra el análisis en que se agrupan las partes de la oración en sintagmas y núcleos, mientras en la figura 3-3b los elementos están ordenados según sus dependencias: el verbo principal de la oración toma la posición superior y las funciones complementarias relacionadas con el verbo se colocan en posiciones subordinadas.

3.2. Aprendizaje máquina

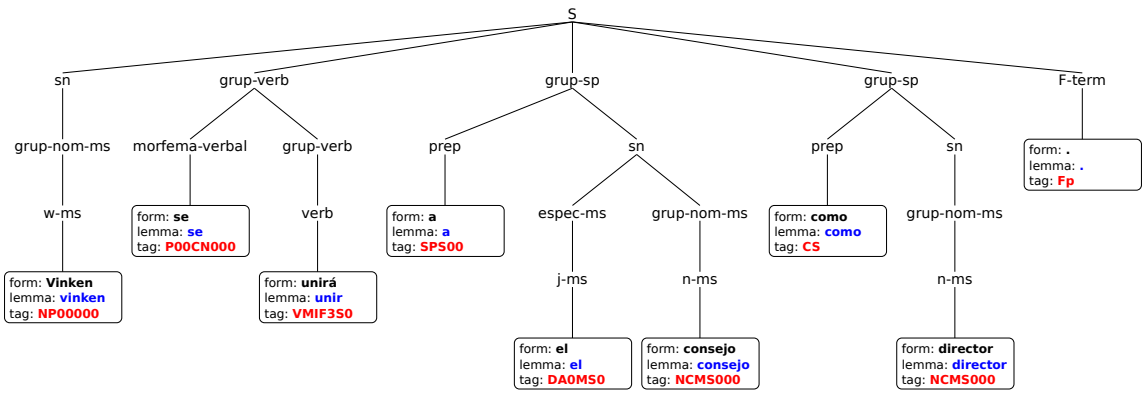
El análisis del lenguaje requiere del razonamiento de un experto del dominio, así como estudios para encontrar patrones y reglas que se puedan seguir, a fin de replicar en una computadora las decisiones del humano. Para ello se puede utilizar aprendizaje automático.

Los trabajos de etiquetado automático se logran gracias al entrenamiento de las máquinas sobre un conjunto de documentos, o corpus, previamente etiquetados. Los corpus son etiquetados por personas que siguen el comportamiento que se busca obtener de la computadora.

3.2.1. Aprendizaje supervisado

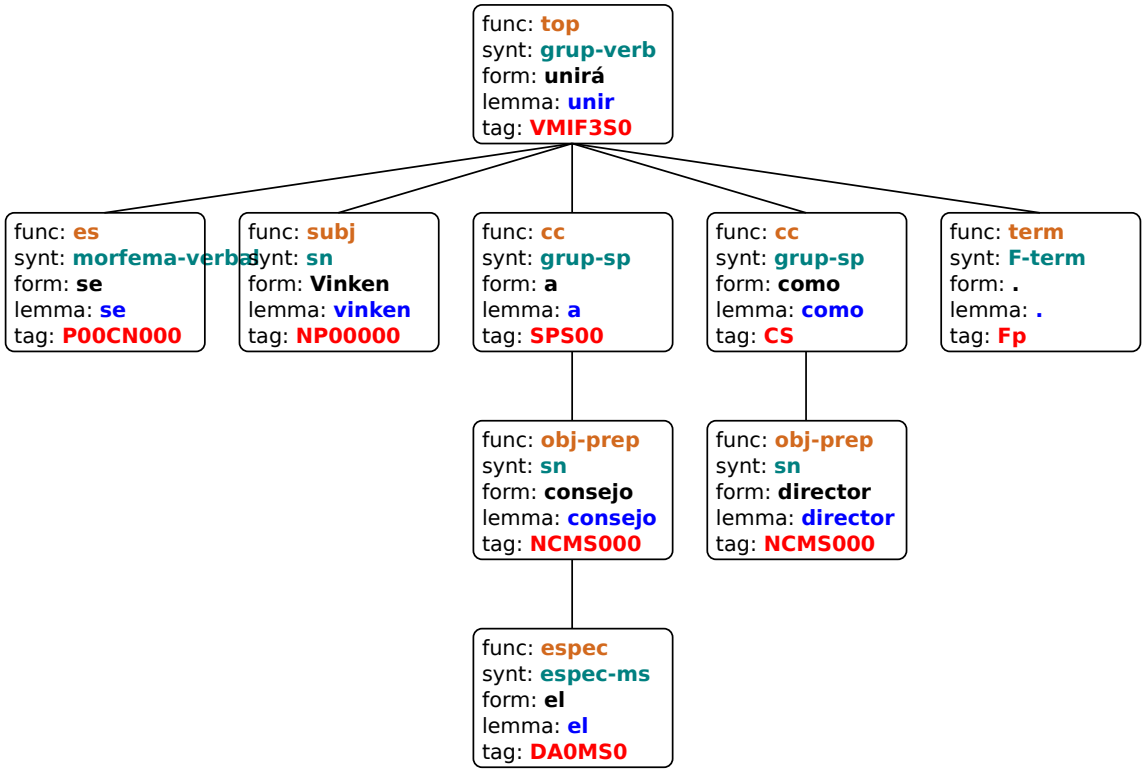
El aprendizaje supervisado es el método de aprendizaje automático mediante el cual una computadora infiere una función a partir de datos anotados, es decir, datos de entrada para los cuales la salida es conocida. Se tiene entonces un conjunto de datos de entrenamiento, pares compuestos por una entrada representada en forma de vector y una salida deseada. El algoritmo entonces analiza los datos de entrenamiento para inferir una función que le permita extrapolar el comportamiento aprendido para casos que no necesariamente haya recibido antes.

El desempeño de estos sistemas depende de muchos factores. Los resultados de una función sencilla serán mucho más fáciles de predecir que los de una función compleja (como funciones con muchas interacciones entre una gran cantidad de variables). Entre menor sea la complejidad de una función, será necesaria una menor cantidad de ejemplos anotados para lograr un aprendizaje adecuado. Sin embargo, conforme la complejidad aumenta, la cantidad de datos para el entrenamiento también debe aumentar. La clasificación de propiedades lingüísticas sigue una gran cantidad de reglas, con numerosas excepciones en donde se involucran muchos factores que



(a) Análisis sintáctico automático

Una oración se puede analizar sintácticamente de forma automática para obtener la división de la oración en sus diferentes componentes estructurales



(b) Análisis de dependencias automático

Una oración se puede analizar sintácticamente con gramáticas de dependencia para obtener la función de las palabras de forma jerárquica dentro de la oración.

Figura 3-3: Comparación entre análisis sintácticos

Comparación del análisis sintáctico de la oración: *Vinken se unirá al consejo como director*. Se muestran el orden estructural y de dependencias de la oración.

se deben de considerar. La complejidad de las funciones lingüísticas restringen el entrenamiento de sistemas a áreas especificas de estudio según sea el interés particular del sistema, y obliga también a utilizar grandes corpus especializados para dichas áreas Jurafsky y Martin [2007].

3.2.2. Aprendizaje no supervisado

A diferencia del aprendizaje supervisado, con el aprendizaje no supervisado no se plantea el uso de datos previamente anotados para lograr las salidas deseadas. Al no tener una guía externa para predecir y generalizar un conocimiento *a priori* de las categorías de salida, la forma en la que se trabaja es mediante agrupación de entradas con características similares en categorías creadas únicamente según la distribución de los datos, este proceso es llamado *clustering*.

De estos métodos de agrupamiento, uno de los más utilizados es el *k-means*. Tal procedimiento utiliza una relocalización iterativa de objetos en k categorías, cada iteración minimiza la distorsión entre el conjunto de objetos y los k *clusters*. Los *clusters* son representados por el punto centroide de los elementos que conforman el *cluster*, el centroide se recalcula y se repite el proceso hasta un criterio de cierre [Duda *et al.*, 2012].

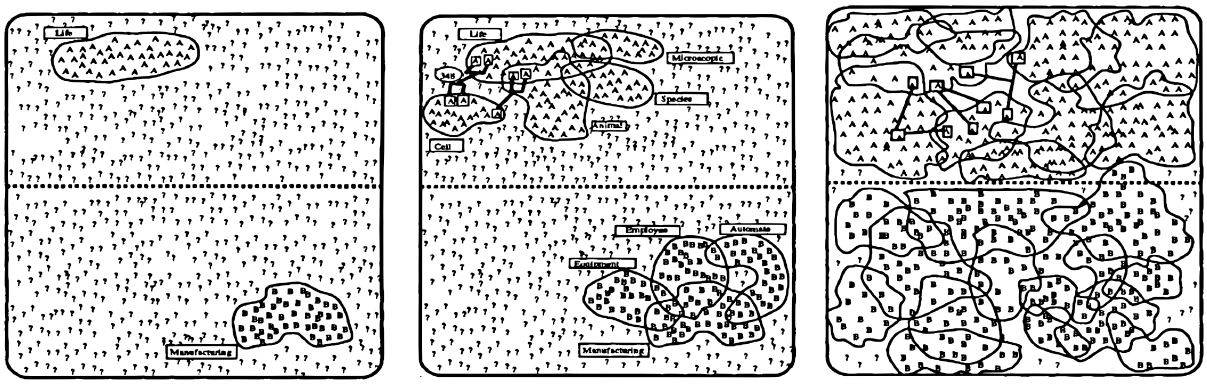
3.2.3. Aprendizaje semi-supervisado

Los dos tipos de aprendizaje antes visto tienen sus ventajas y desventajas. Por un lado, en el aprendizaje supervisado se pueden obtener mejores resultados al entrenar a la máquina con las categorías que uno necesita, aunque se tiene la desventaja de que el método necesita grandes cantidades de datos anotados. Por otro lado, se tiene el aprendizaje no supervisado que es capaz de trabajar sin los datos previamente clasificados para encontrar grupos de elementos similares entre sí.

En un punto medio, se encuentra el aprendizaje semi-supervisado. La importancia de este método radica en su capacidad para trabajar con conjuntos grandes de datos, de los cuales sólo una pequeña parte debe estar clasificada con la categorización deseada [Zhou *et al.*, 2004].

En la figura 3-4 se muestra un ejemplo del funcionamiento de este tipo de entrenamiento. El sistema comienza con un conjunto de objetos clasificados como A y B dentro de un corpus, donde la mayoría de los objetos a clasificar no cuentan con una etiqueta. El algoritmo iden-

tifica los objetos más cercanos a los elementos anotados de entrenamiento y los agrega a la misma categoría. En la siguiente iteración, toma todos los nuevos elementos etiquetados como el conjunto de entrenamiento y el algoritmo se repite hasta alcanzar un criterio final.



(a) (b) (c)

Figura 3-4: Entrenamiento semi-supervisado [Jurafsky y Martin, 2007]

Proceso de aprendizaje semi-supervisado. Inicia con un conjunto de entrenamiento pequeño. Obtiene los elementos más cercanos y los agrega al entrenamiento. Finalmente, repite la operación hasta alcanzar un objetivo.

Imagen tomada de *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* [Jurafsky y Martin, 2007].

Dentro del aprendizaje semi-supervisado, existe una caso especial llamado aprendizaje activo, que se caracteriza por permitir al sistema preguntar al usuario sobre nuevas etiquetas para anexar al corpus de entrenamiento. Este método es principalmente efectivo si se utiliza para etiquetar los elementos que quedan más lejos de los centroides originales, es decir, cuando se agregan al corpus de entrenamiento los elementos de mayor incertidumbre [Seeger, 2000].

3.3. Métricas de evaluación

Los resultados de un sistema automático aportan muy poca información si se desconoce qué tan efectivo es el sistema para la tarea que está desempeñando. Existen diferentes técnicas convencionales para medir el desempeño de los sistemas de clasificación. En esta sección se describen tres métricas comúnmente utilizadas para la evaluación de sistemas de PLN [van Rijsbergen, 1995].

3.3.1. Precisión

La precisión es el radio que existe entre los elementos que fueron recuperados correctamente dentro del total de los elementos recuperados [van Rijsbergen, 1995].

Sea por ejemplo un conjunto de elementos dentro de los cuales se encuentran los elementos de interés que quieren ser recuperados o etiquetados con cierta categoría, y una clasificación automática para lograr la tarea. La precisión mide, dentro de todos los elementos que fueron extraídos o etiquetados con la categoría de interés, qué porcentaje fue logrado con éxito. En la figura 3-5a se muestra un esquema de conjuntos en donde se explica la precisión de forma gráfica.

3.3.2. Exhaustividad

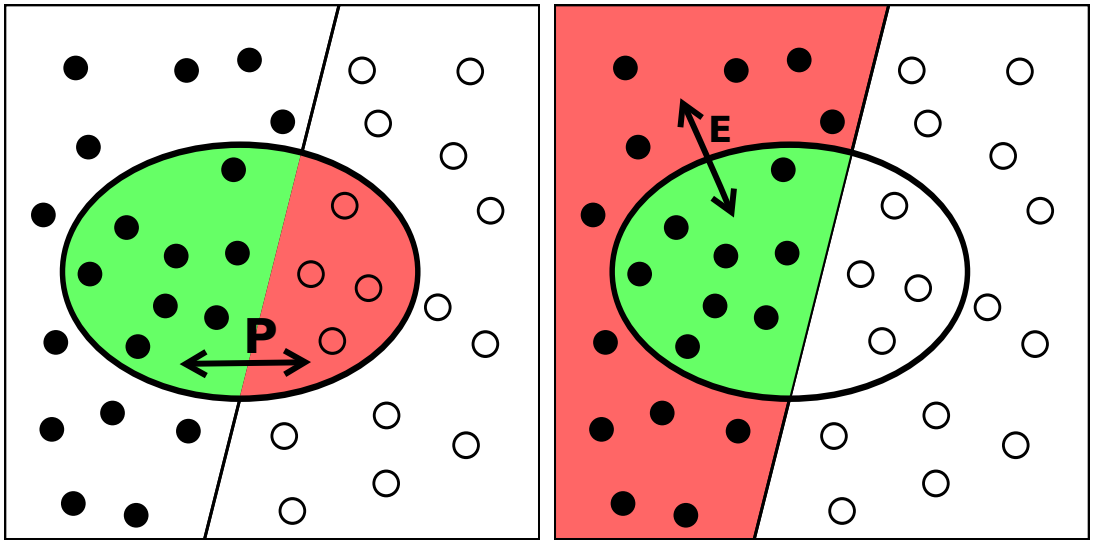
La exhaustividad es la razón que existe entre los elementos que fueron recuperados correctamente dentro del total de los elementos que se deberían recuperar por un sistema.

Considérese que se tiene un conjunto de elementos dentro de los cuales se encuentran los elementos de interés que quieren ser recuperados o etiquetados con cierta categoría, y una clasificación automática para lograr la tarea. La exhaustividad mide, dentro de todos los elementos que debieron ser extraídos o etiquetados con la categoría de interés, qué porcentaje fue etiquetado con éxito. En la figura 3-5b se muestra un esquema de conjuntos en donde se explica la exhaustividad de forma gráfica.

3.3.3. Medida-F

La medida-F es un valor que mide el desempeño general que tiene la salida de un sistema. Se trata de un valor único que toma en cuenta tanto la precisión como la exhaustividad, de tal forma que:

$$F = 2 \cdot \frac{Precisión \cdot Exhaustividad}{Precisión + Exhaustividad}$$



(a) Precisión

La precisión es la relación que existe entre los elementos correctamente extraídos con respecto al total de los extraídos.

(b) Exhaustividad

La exhaustividad es la relación que existe entre los elementos correctamente extraídos con respecto al total de los que se buscaban.

Figura 3-5: Métricas de evaluación de precisión y exhaustividad

En las figuras se muestran en negro los elementos correctos a ser recuperados, el óvalo contiene los elementos extraídos por el sistema a evaluar. Las zonas verdes representan aciertos, mientras que las rojas representan errores.

3.3.4. Validación cruzada

En adición a las métricas antes mencionadas, existen técnicas que permiten una mayor confianza en el momento de la evaluación, un ejemplo de esto es la validación cruzada. La validación cruzada es una técnica que consiste en dividir los datos que se tienen para la evaluación en partes iguales. Cada una de las partes, de manera iterativa, se toma para evaluar mientras el resto se toma para entrenar. Al final del proceso, se toma la media de las evaluaciones. En la figura 3-6 se muestra un esquema de validación cruzada de cuatro divisiones.

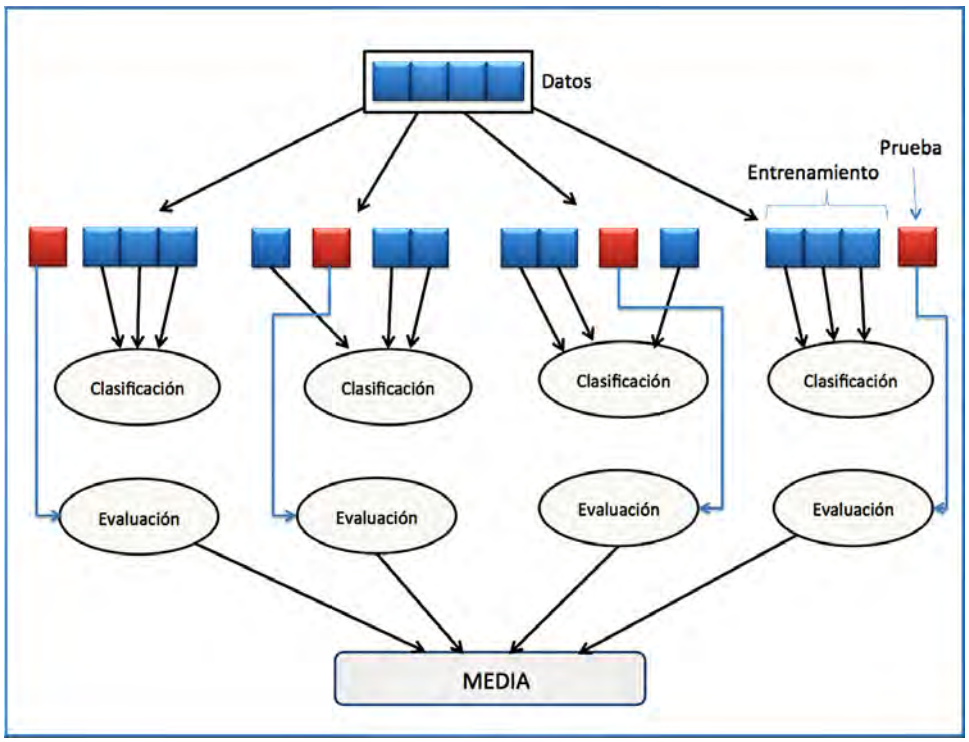


Figura 3-6: Validación cruzada

Los datos anotados son divididos en partes iguales (para este caso 4 partes) y se hace un entrenamiento tomando cada parte por separado como prueba. En la figura, las partes azules representan los datos que se usan como entrenamiento, la parte roja es la utilizada como prueba, se hacen igual número de evaluaciones y de divisiones y se obtiene una media final.

Imagen por Joan.domenech91 [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

3.4. Relaciones temporales

Dentro del desarrollo de esta tesis, el manejo de la información temporal se realiza utilizando el enfoque de búsqueda de relaciones temporales entre eventos. Esto quiere decir que dentro del texto se encuentra información que establece, o da indicios de la forma en la que dos eventos se pueden relacionar, por ejemplo, que un evento haya ocurrido antes, durante o después de otro evento. Las relaciones temporales pueden ser consideradas de varias maneras, y la cantidad de relaciones a tomar en cuenta también es variable.

3.4.1. Temporalidad lingüística

Dentro de la lingüística, Rojo [1990] describe la temporalidad lingüística como una categoría gramatical deíctica mediante la cual se expresa la orientación temporal de una situación. La orientación temporal se entiende como una de tres relaciones: anterioridad, simultaneidad o posterioridad. El tiempo de creación del documento se convierte en punto cero o inicial con respecto al cual se relacionan los primeros eventos. A partir de eventos que ya tienen una dirección temporal se pueden relacionar nuevos hechos, lo que da lugar a relaciones temporales más complejas en el encadenamiento de escalones de eventos. La jerarquía que propone Rojo [1990] se expone en el diagrama 3-7, en donde *A* es un punto que puede ser definido simplemente como anterior al origen, pero *S'* es presentado como simultáneo al punto anterior al origen. Todos los ejemplos que se manejan en dicho trabajo muestran la construcción de relaciones temporales entre verbos y verbos subordinados dentro de oraciones compuestas.

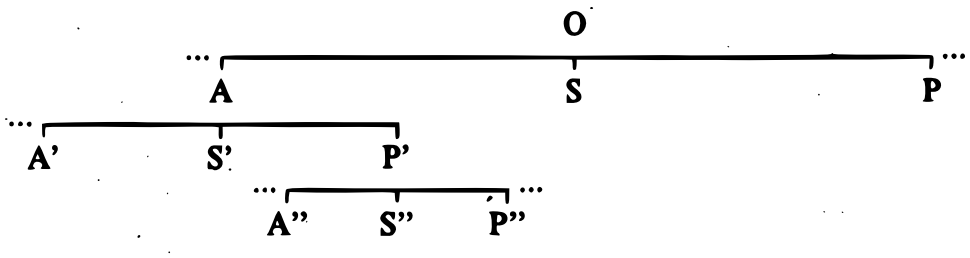


Figura 3-7: Diagrama de temporalidad lingüística [Rojo, 1990]

La complejidad de las relaciones temporales surge de un encadenamiento de eventos que mantienen relaciones simples.

Diagrama tomado de *Relaciones entre temporalidad y aspecto en el verbo español* [Rojo, 1990].

3.4.2. Relaciones de Allen

Desde un punto de vista más computacional, Allen [1983] afirma que un buen modelo y representación del tiempo debe cumplir con cuatro puntos principales:

- Debe permitir imprecisión. Mucho conocimiento temporal es relativo (por ejemplo *"A fue antes que B"*) y tiene poca relación con fechas absolutas.

- Debe permitir incertidumbre en la información. En muchas ocasiones, la relación exacta entre dos tiempos es desconocida, pero se pueden obtener pistas y acotaciones acerca del cómo pueden estar relacionados. La transitividad presente en el álgebra desarrollada por el mismo Allen (abordadas más adelante en esta sección, en la figura 3-9 y en el ejemplo 3.4.1) es un claro ejemplo de incertidumbre.

- Debe permitir variación en la dimensión del razonamiento. En el caso de historia, por ejemplo, se puede considerar el tiempo sólo en término de días o incluso años. Mientras que en el manejo de fenómenos o diseños, es necesario considerar dimensiones como milisegundos o menos.

- Debe permitir persistencia. Es decir, debe ser capaz de considerar razonamientos como *"Si dejé el coche estacionado afuera esta mañana, debería estar todavía allí ahora"*, a pesar de que no sea posible la comprobación del razonamiento (el carro pudo ser robado).

Además de eso, Allen [1984] hace un estudio para establecer las diferentes posibles relaciones temporales que pueden existir entre un par de eventos. En la figura 3-8 se muestran las trece relaciones que pueden existir entre dos eventos. La tabla muestra tres columnas:

- La columna de la izquierda señala la notación que se tiene para cada una de las relaciones. Las relaciones se encuentran emparejadas con las relaciones inversas, de manera que en cada renglón de esta columna se pueden observar dos símbolos, uno que relaciona al evento X con el evento Y, y un segundo, que al ser inverso del primero, relaciona de manera contraria a Y con X.

- La columna central muestra de manera gráfica la relación. Todas las relaciones están dadas para un par de eventos X y Y. En la ilustración se muestra una línea que representa de

manera esquemática el tiempo de duración de los eventos. La posición de las líneas entre sí expresa la relación que se quiere mostrar. Para cada una de las ilustraciones corresponden dos relaciones que son inversas, con excepción de la relación **igual**, cuyo inverso es la misma relación.

- La columna de la derecha contiene la interpretación de la relación escrita con palabras.

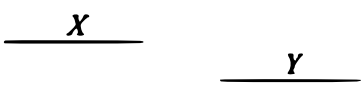
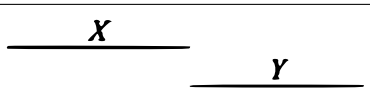
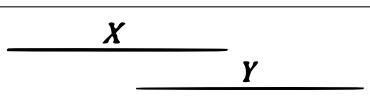
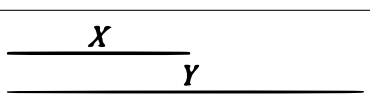
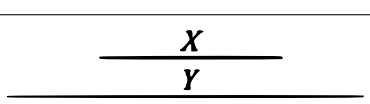
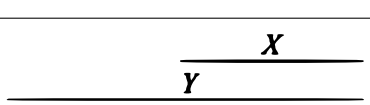
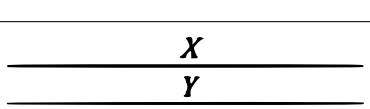
Relación	**Ilustración**	**Interpretación**
$X<Y$ $Y>X$		X ocurre antes que Y
X **m** Y Y **mi** X		X ocurre justo antes que Y
X **o** Y Y **oi** X		X se traslapa con Y
X **s** Y Y **si** X		X comienza a Y
X **d** Y Y **di** X		X durante Y
X **f** Y Y **fi** X		X termina a Y
$X=Y$		X igual a Y

Figura 3-8: Relaciones temporales de Allen [1983]

La figura muestra una tabla en la que se ejemplifican las trece relaciones temporales introducidas por James F. Allen en 1983, de forma gráfica se representa la duración de cada par de eventos (x e y) con una línea, la relación se observa según la posición de un intervalo de duración de un evento con respecto al otro

Ilustraciones tomadas de *https://en.wikipedia.org/wiki/Allen%27s_interval_algebra*

Las trece relaciones de Allen [1984] son fundamentales para los procesos de esta tesis. Son dichas trece categorías las que se toman en cuenta para vincular todos los pares de eventos que se manejan.

Una aspecto importante con el que cuenta el trabajo de Allen [1983] es que al tener un evento relacionado con un segundo evento, y este último se relaciona con un tercero, se pueden relacionar también el primer evento con el tercero. Dicho autor estudia este fenómeno y desarrolla un álgebra de intervalos para las trece categorías que permite conocer la forma en la que se vinculan los eventos que se encuentren indirectamente relacionados. Para representar la incertidumbre de la relación concreta entre un par de eventos utiliza operaciones lógicas con las relaciones posibles entre dicho par. En la figura 3-9 se muestran las operaciones de transitividad en el álgebra desarrollada por Allen [1983]. En la columna de la izquierda se muestra la primera relación temporal, la que es del evento A con el evento B. En el primer renglón, se muestra la segunda relación, es decir la que correspondería a la del evento B con un tercer evento C. La intersección de dos relaciones representa las disyunciones entre las posibles relaciones que se puede presentar entre el primer evento A y el tercer evento C.

Ejemplo 3.4.1 Como ejemplo práctico, podemos hablar de la relación temporal entre Miguel Hidalgo, el grito de Dolores y con la independencia de México.

- Tomaremos la vida de Miguel Hidalgo como el evento A
- Tomaremos el grito de Dolores como el evento B
- Tomaremos la Guerra de Independencia como el evento C

Consideremos también dos relaciones dadas:

- El grito de Dolores transcurre dentro de la vida de Miguel Hidalgo, es decir: A incluye B. Siguiendo la notación de la figura 3-8, tendríamos que A *di* B
- El grito de Dolores da inicio a la guerra de independencia de México, es decir: B comienza a C. Siguiendo la notación de la figura 3-8, tendríamos que B *s* C

Con la información de las dos relaciones anteriores, podemos buscar en la tabla de transitividad. En la columna izquierda buscamos la primer relación de A con B, es decir, buscamos *di*, que se encuentra en la cuarta posición. Para la segunda relación, la de B con C, vamos al primer renglón en busca de la relación *s*, que se encuentra en la novena columna.

La intersección de la columna y el renglón encontrados nos da las posibles relaciones que existen entre A y B (entre la vida de Miguel Hidalgo y la guerra de independencia de México)

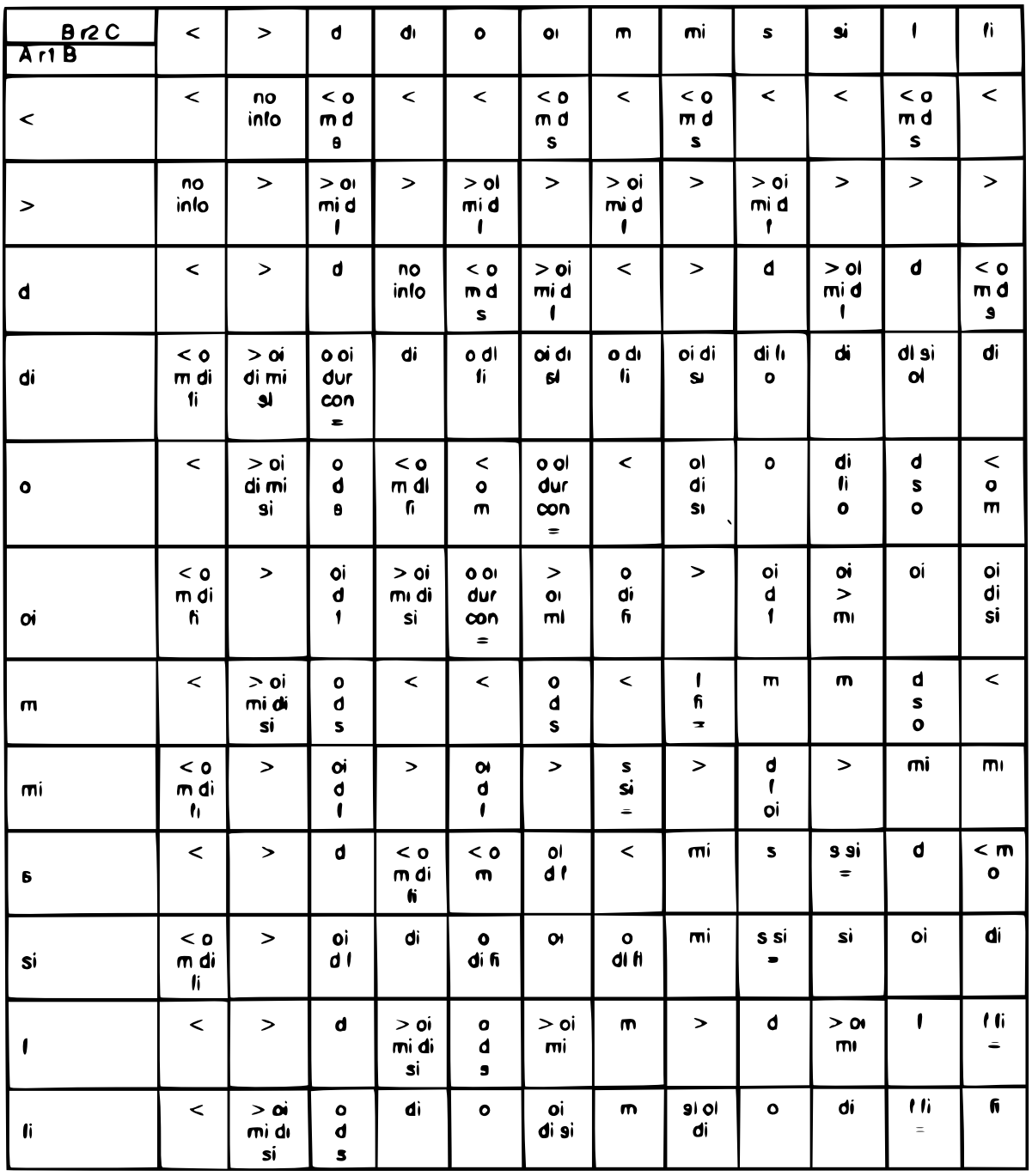


Figura 3-9: Álgebra de intervalos temporales [Allen, 1983]

Transitividad entre relaciones temporales [Allen, 1983]. Si se tiene un evento A relacionado con B con la relación de la columna izquierda, y se tiene el evento B relacionado con C según el primer renglón, entonces los eventos A y C tendrán alguna de las relaciones acotadas por la intersección de la tabla.

Tabla tomada de *Maintaining knowledge about temporal intervals* [Allen, 1983]

tan solo con la información de los puntos anteriores. Para el caso, se encuentran las relaciones: *di fi* y *o* como las tres posibles relaciones temporales entre A y B:

- Puede ser que la vida de Miguel Hidalgo incluya a la guerra de independencia.
- O bien, puede ser que la vida de Miguel Hidalgo termine a la independencia.
- O bien, puede ser que la vida de Miguel Hidalgo se traslape con la independencia

La información incompleta que se consideró desde un principio da lugar a incertidumbre, con información extra se pueden hacer más reducciones hasta llegar a la opción correcta.

3.4.3. Vecindades de Freksa

El álgebra de intervalos propuesta por Allen ha demostrado tener problemas de complejidad al trabajar con grandes cantidades de relaciones indirectas para resolver. Esto ha dado lugar a nuevas propuestas para manejar la información de las relaciones, y otros enfoques como el álgebra de puntos temporales [Vilain y Kautz, 1986].

Freksa [1992] retoma el uso de intervalos temporales para desarrollar su trabajo con una perspectiva cognitiva. Toma en cuenta vecindarios de relaciones temporales vinculados por restricciones de percepción. En la figura 3-10 se muestra la representación esquemática de Freksa [1992] para las trece relaciones de Allen [1983]. Dentro de la figura se pueden apreciar dos estructuras, una del lado izquierdo y la otra del lado derecho; ambas construcciones contienen las trece relaciones de Allen [1983] de manera ordenada. Del lado derecho se muestran las relaciones representadas por la notación que se introdujo en la sección 3.4.2 dentro de la figura 3-8; cada símbolo de relación introducido en la figura antes mencionada se encuentra dentro de un círculo. Del lado izquierdo, para cada una de las relaciones del lado derecho, se muestran una representación esquemática de los intervalos de dos eventos que participan en dicha relación temporal, este tipo de representación es introducida por Freksa [1992] y se explica a continuación.

Cada uno de los esquemas presentes en el lado izquierdo de la figura 3-10 contiene la representación de la duración de un evento hipotético, el primero de los cuales se representa mediante dos círculos unidos por una línea (❍❍), y el segundo se representa con un rectángulo (❑). De esta manera, si en la estructura de la derecha se tiene la relación $<$, del lado izquierdo se tendrá la representación de ❍❍ $<$ ❑ al mostrar, como se puede observar en la figura, el intervalo

representado por los dos círculos del lado izquierdo (antes) del rectángulo. La posición relativa de los extremos de los eventos representados es lo que muestra la relación que existe entre ellos.

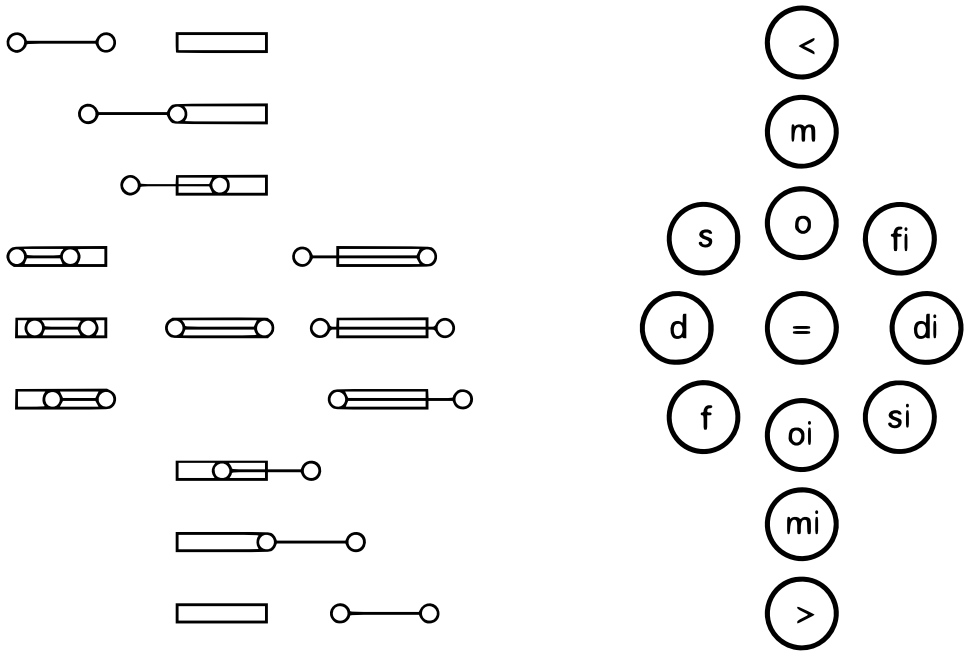


Figura 3-10: Estructura de vecindades de relaciones temporales [Freksa, 1992]

Estructura propuesta por Freksa para agrupar las relaciones temporales en vecindades conceptuales, del lado izquierdo se muestran las relaciones esquemáticas y del lado derecho sus símbolos correspondientes.

Diagrama tomado de *Temporal reasoning based on semi-intervals* [Freksa, 1992].

El objetivo de esta nueva forma de ordenar las relaciones temporales es mantener una representación del tiempo simple a pesar de trabajar con información incompleta sobre las relaciones. Las relaciones temporales pueden ganar dependencia al quedar agrupadas según la posibilidad de transformar directamente una relación en otra contigua, a continuación se explica esto con más detalle.

Se tomará como apoyo nuevamente la figura 3-10. En la parte izquierda de la figura, se muestran esquemáticamente las trece relaciones temporales; es conveniente notar que en cada par de intervalo representado (❍❍ y ❑), el evento representado con el rectángulo (❑) se "queda quieto"mientras el evento representado con los círculos y la línea (❍❍) se "mueve"horizontalmente al ir cambiando de relación temporal. Este comportamiento es el que observa Freksa [1992] y con el que describe las relaciones que se pueden transformar entre si. Según lo establece, se conectan aquellas relaciones que al mover uno solo de los extremos de uno de los dos intervalos

se llega a otra relación. En la figura 3-10 podemos ver esta transformación de una relación en otra, por ejemplo, a partir de la primer relación (❍❍ < ❑) podemos llegar a la segunda (❍❍ m ❑) si mantenemos el evento rectangular (❑) fijo y movemos el extremo derecho del evento representado con círculos (❍❍) a la derecha hasta que llegue al extremo izquierdo del evento rectangular (❑). A partir de la primera relación (❍❍ < ❑) no se puede llegar a ninguna otra al mover únicamente un extremo de uno de los eventos, por lo tanto, la primera relación solo se conecta con la segunda (❍❍ m ❑). A diferencia de lo anterior, si tomamos la segunda relación como referencia (❍❍ m ❑), se puede transformar al mover solo un extremo tanto en la primera ((❍❍ < ❑)) como en la tercera (❍❍ o ❑); y si tomamos como referencia la tercera relación (❍❍ o ❑), se puede transformar directamente en las relaciones: {❍❍ **m** ❑}, {❍❍ **s** ❑} y {❍❍ **fi** ❑}.

En la figura 3-11 se muestran las conexiones que considera relaciones adyacentes a las que se pueden transformar directamente, es decir, si se puede convertir una relación temporal en otra al mover uno de los extremos de la relación. Dentro de la figura se muestran conectadas por una línea las relaciones que presentan esta propiedad; y son estas mismas conexiones las que le dan la forma al diagrama con el que Freksa [1992] presenta las relaciones temporales.

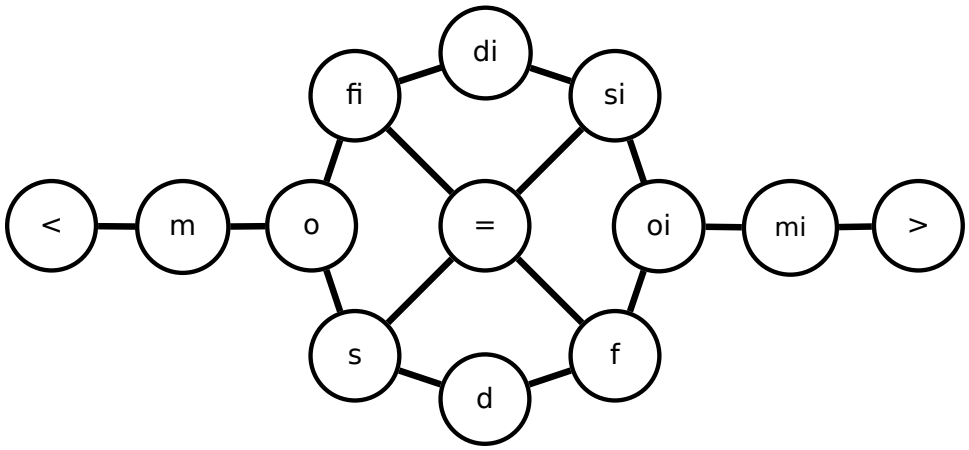


Figura 3-11: Conexión entre relaciones temporales adyacentes [Freksa, 1992]

Las relaciones temporales pueden ser ordenadas al poner de forma contigua a las relaciones que permiten una transformación directa entre sí. En la figura se muestran las trece relaciones temporales conectadas con aquéllas con las que una transformación de este tipo es posible [Freksa, 1992].

Diagrama extraído de *Temporal reasoning based on semi-intervals* [Freksa, 1992].

Dentro de la estructura, un conjunto de relaciones que se mantenga unida a través de las conexiones mostradas, se considera una vecindad de relaciones temporales. Las vecindades pueden ser utilizadas para manejar la información temporal de una manera más simple en los casos de incertidumbre. El álgebra de Allen [1983], como se observa en la figura 3-9, tiene como resultado la disyunción de un grupo de relaciones temporales. Freksa hace notar que dichos grupos de relaciones forman vecindades dentro de su diagrama. En la figura 3-12 se muestra nuevamente el álgebra de Allen, esta vez representando las relaciones con el diagrama de Freksa. Dentro de la figura, se observa nuevamente la tabla de transitividad de la figura 3-9 con un cambio de simbología; el diagrama de las figuras 3-10 y 3-11 se presenta de forma esquemática, denotando únicamente con un punto negro la ubicación de la relación temporal a la que se está refiriendo.

Freksa hace una serie de observaciones importantes con respecto a los resultados de la transitividad. Entre los más importantes están el notar que todos los resultados son puntos continuos en el diagrama y, además, son una pequeña fracción de las posibles combinaciones que se pueden lograr. Más adelante aplica la transitividad con dicho conjunto de salidas hasta lograr un conjunto cerrado en el que participan únicamente 29 expresiones.

Por otra parte, reconoce 18 conjuntos de relaciones que aparecen frecuentemente. Dichas vecindades de relaciones son comunes de encontrar, tanto por el resultado del álgebra temporal como por posibles interpretaciones de una relación, en el caso de que un texto no identifique claramente la relación temporal que existe entre dos eventos narrados. En la figura 3-13 se pueden ver los 18 vecindarios que obtiene Freksa.

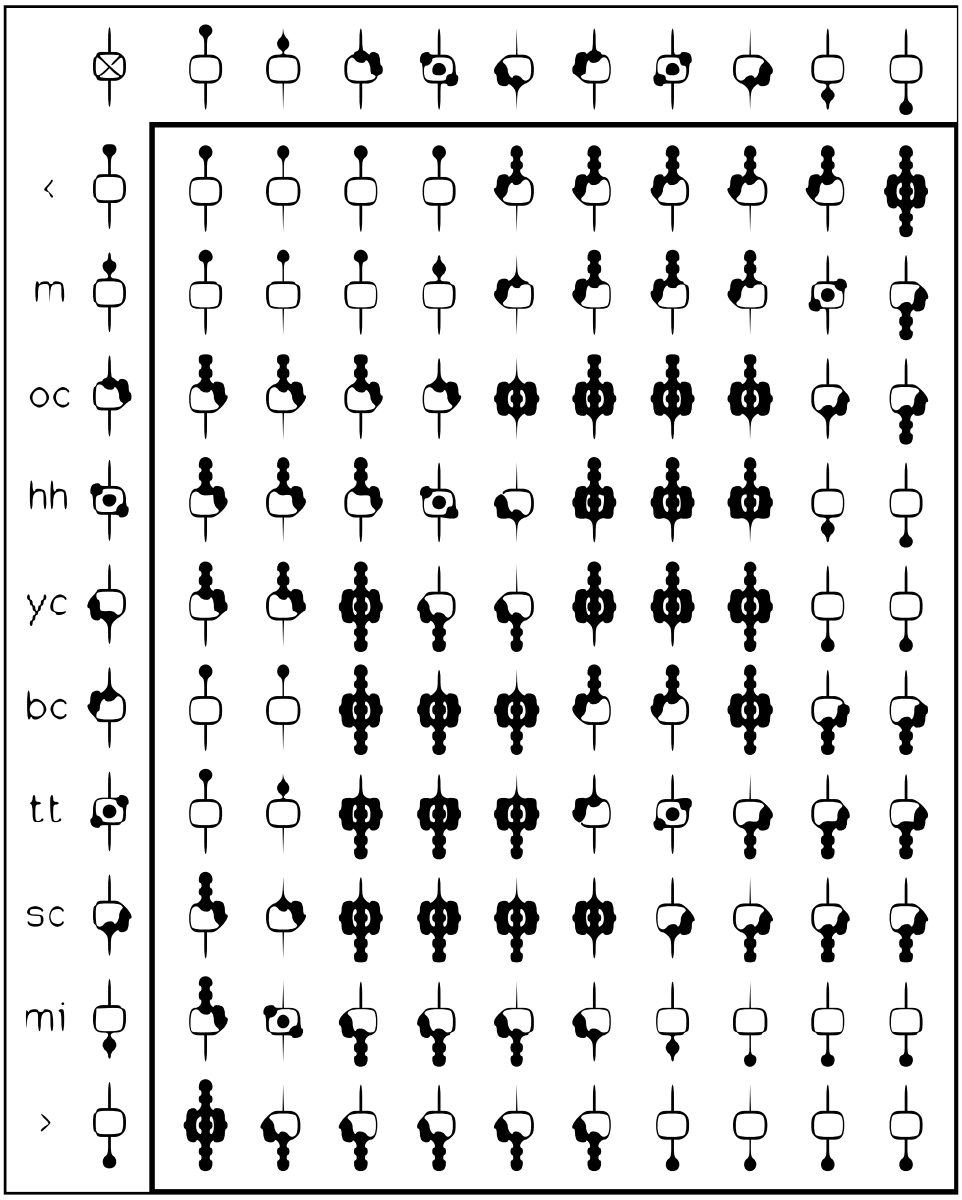


Figura 3-12: Álgebra de Allen [1983] en representación de [Freksa, 1992]

Transitividad entre relaciones temporales expresada con la simbología de vecindades conceptuales.

Tabla tomada de *Temporal reasoning based on semi-intervals* [Freksa, 1992].

Icono	Significado
	sin información
	más viejo que
	inicio con inicio
	más joven que
	es sobrevivido por
	final con final
	sobrevive a
	precede a
	nace antes de la muerte de
	contemporáneo con
	muere después del nacimiento de
	sucede a
	más viejo y sobrevivido por
	más viejo y contemporáneo con
	contemporáneo y sobrevive a
	contemporáneo y sobrevivido por
	más joven y contemporáneo con
	más joven y sobrevive a

Figura 3-13: Extensión de relaciones temporales [Freksa, 1992]

Iconos e interpretaciones de los vecindarios temporales más comunes hallados por Christian Freksa [1992].

Tabla tomada de *Temporal reasoning based on semi-intervals* [Freksa, 1992].

Capítulo 4

Metodología

Para esta tesis se utiliza un enfoque de aprendizaje semi-supervisado con el fin de maximizar el desempeño del sistema. El desarrollo del sistema ordenador consiste en tres etapas principales esquematizadas en la figura 4-1.

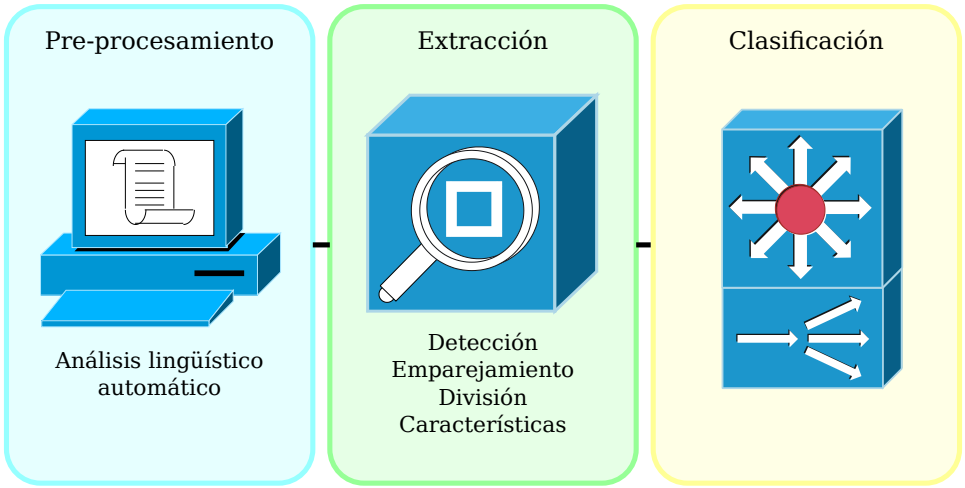


Figura 4-1: Metodología
El desarrollo del sistema consta de un proceso de tres etapas principales.

En primer lugar, se tiene el pre-procesamiento. Esta etapa consiste en el análisis sintáctico de dependencias de los textos que se usan como corpus de entrenamiento.

Tras expresar las oraciones de los textos como árboles de dependencias, el siguiente paso es procesarlos para obtener la información de interés. Se extraen eventos, conexiones que emparejan los eventos y las características que rodean dicha conexión.

El tercer bloque se encarga de la clasificación, consta de dos procesos importantes. En primer lugar se realiza una clasificación gruesa de relación temporal para cada pareja de eventos. La clasificación se hace entre un conjunto de relaciones basado en las vecindades de Freksa. De esta clasificación se obtiene una primera aproximación de las relaciones temporales. El segundo y último bloque de la clasificación y del proceso realiza la clasificación fina para obtener las etiquetas de las relaciones según la representación de Freksa [1992]. El resultado obtenido de esta etapa representa el ordenamiento final.

4.1. Pre-procesamiento

Los documentos de entrada son sometidos a un pre-procesamiento para obtener la información lingüística básica de los textos de forma automática. Como se muestra en la figura 4-2 de manera esquemática, la información obtenida consta de un árbol sintáctico que contiene las etiquetas de los análisis lingüísticos automáticos para cada una de las oraciones.

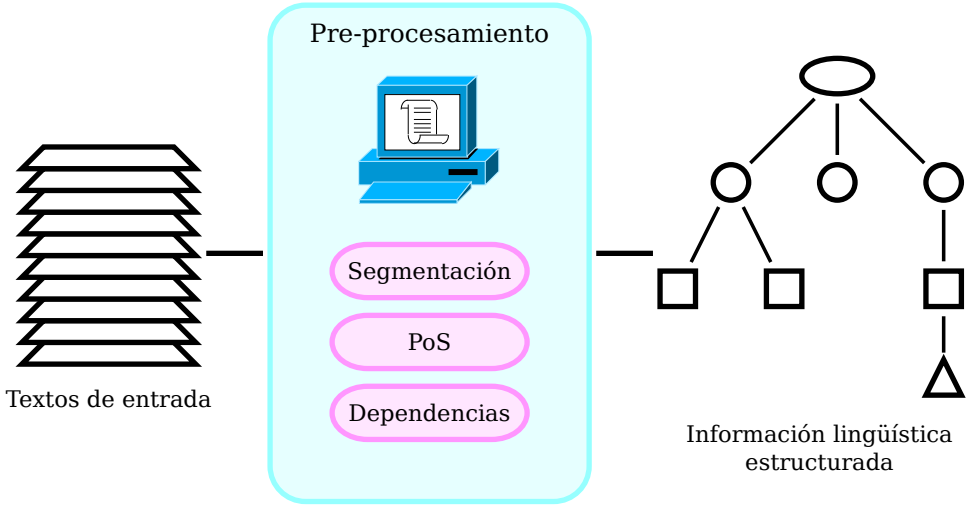


Figura 4-2: Pre-procesamiento

La etapa de pre-procesamiento trabaja con los documentos de entrada para obtener información lingüística de las oraciones ordenada dentro de un árbol sintáctico.

El analizador automático que se utiliza para este trabajo se encarga de tres procesos principales. En primer lugar, separa el texto en oraciones y las oraciones en palabras o segmentos. Posteriormente, analiza la morfología de las palabras, les asigna sus categorías y realiza un etiquetado *PoS*. Por último, se ejecuta el análisis sintáctico para encontrar las dependencias

sintácticas dentro del texto.

Del análisis lingüístico automático se obtiene, para cada archivo de texto de entrada, un conjunto de estructuras, cada una de las cuales formada por un árbol de dependencia para cada oración del texto.

A partir de dichos árboles se extraen los eventos a considerar, los pares de eventos que posiblemente guardan alguna relación temporal y las características que el sistema utiliza para el momento de hacer la clasificación de relaciones.

4.2. Extracción de información

El concepto general de esta etapa del procesamiento se resume esquemáticamente dentro de la figura 4-3. La etapa de extracción consta de cuatro partes. En primer lugar, se identifican todos los eventos a partir de la información lingüística que se obtuvo en la etapa de pre-procesamiento. Tomando en cuenta que las relaciones temporales se presentan entre un par de eventos, los eventos encontrados son emparejados en la segunda etapa de este módulo. Se consideran diferentes maneras en las que los eventos pueden quedar emparejados, por lo que en la tercera etapa se hace una división que separa a las parejas según el tipo de conexión que tengan los eventos. Cada uno de los tipos de parejas cuenta con diferentes características que se extraen también de la información lingüística que entra al módulo, esta extracción de características es la última etapa del proceso de extracción. Lo que se obtiene son todas las parejas catalogadas por el tipo de emparejamiento que tienen, y acompañadas de sus respectivas características.

4.2.1. Detección de eventos

Chambers *et al.* [2014] mencionan la importancia de encontrar, a manera automática, los pares de eventos sobre los cuales trabajar para encontrar una relación temporal. Aunado a eso, dentro del estudio de temporalidad lingüística de Guillermo Rojo [1990] se plantea la orientación temporal de verbos dentro de oraciones compuestas. En esta tesis se aplica la propuesta de utilizar la relación sintáctica de dependencias como un indicador de relaciones temporales.

Como uno de sus ejemplos, Rojo [1990] utiliza la oración: *Nos dijo que había llegado el día anterior*. En dicho ejemplo se plantea un primer nivel de dirección de anterioridad con respecto

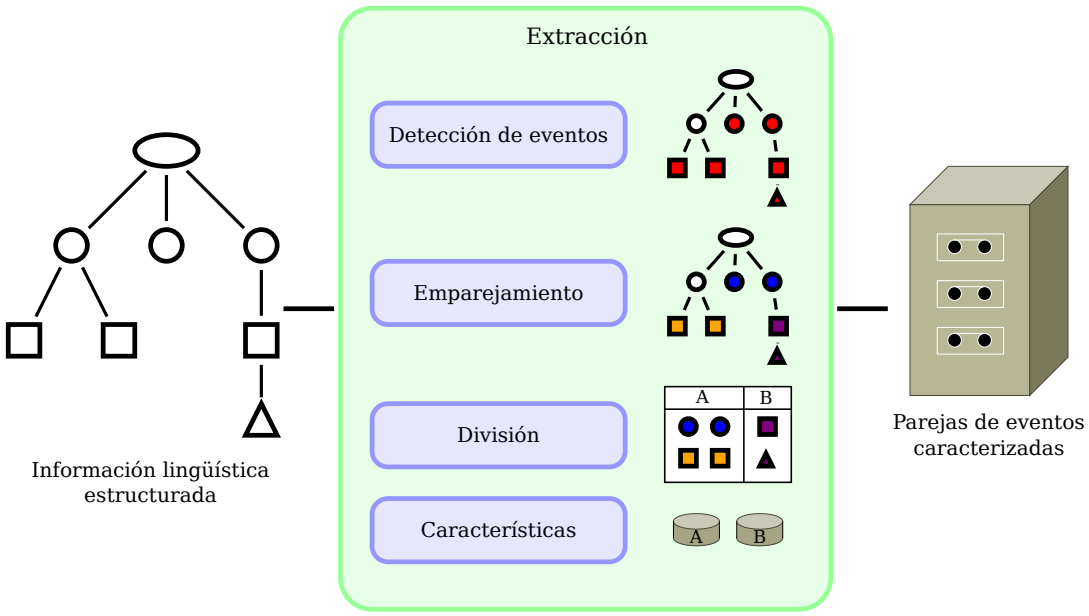


Figura 4-3: Extracción

La etapa de extracción consiste en encontrar los eventos con los que se van a trabajar, y agruparlos en diferentes tipos de parejas. Lo que se obtiene del proceso son las parejas de eventos que se ordenarán junto con las características que serán usadas para el entrenamiento.

al punto de enunciación para la situación *dijo*, sobre la cual se plantea un segundo nivel de dirección de anterioridad para la situación *había llegado*.

Un análisis de dependencias es capaz de mostrar la relación sintáctica que existe entre los verbos principales de las oraciones que constituyen la oración compuesta. En la figura 4-4 se muestra el análisis de dependencias de la oración *Nos dijo que había llegado el día anterior*. Como se muestra en la figura, los verbos centrales de las oraciones se colocan en la posición superior de su respectiva oración y son identificados por el analizador con la etiqueta *'synt:grup-verb'*. Para los procesos de esta tesis, son dichos verbos centrales los que se utilizarán como los eventos dentro del texto.

4.2.2. Emparejamiento y división

Con emparejamiento de eventos nos referimos a la ordenación de los eventos en parejas, entre los dos eventos de cada pareja se buscará una relación temporal. El emparejamiento de eventos y la división de tipos de parejas son procesos vinculados. La división se hace según las diferentes formas en las que se pueden emparejar los eventos, y a partir de eso el emparejamiento

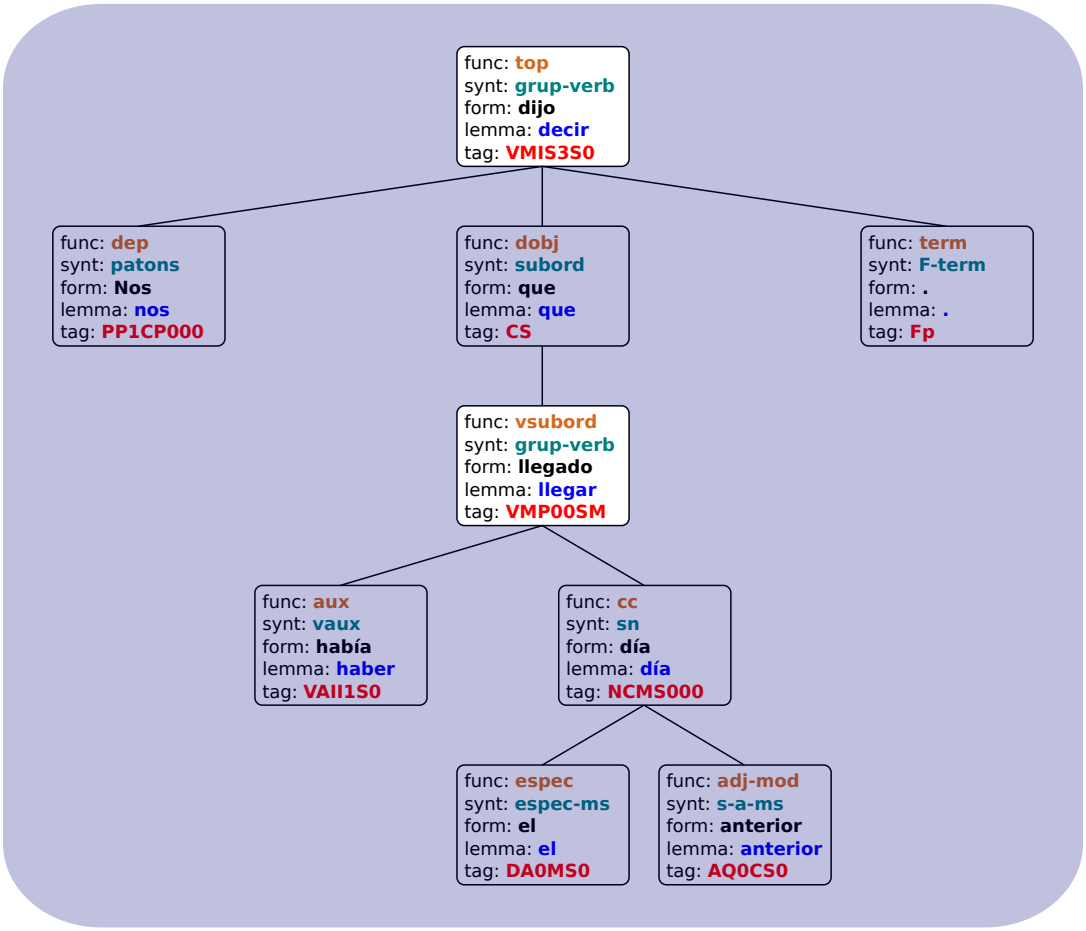


Figura 4-4: Pareja de dependencia

La oración: *Nos dijo que había llegado el día anterior*. Analizada por dependencias, el sistema encuentra dos verbos núcleos en la oración compuesta a los cuales asigna la etiqueta *'synt:grup-verb'*. Dichos verbos se pueden analizar como eventos para encontrar una relación temporal entre el par de *verbos-eventos*.

se hace siguiendo las pautas de las divisiones. Es conveniente pensar en la división como una clasificación del tipo de parejas con las que se trabajará. Nos referimos a este proceso como división para evitar cualquier confusión con los procesos posteriores de clasificación de relaciones temporales. En esta sección se muestra la división de tipos de parejas y el emparejamiento que se toman en cuenta para el desarrollo de este trabajo.

El ejemplo de la figura 4-4 para la oración: *"Nos dijo que había llegado el día anterior"* muestra, dentro de su análisis de dependencias, dos eventos que se pueden emparejar de la forma que consideraremos como dependencia directa, la cual se da entre un verbo principal y un verbo subordinado. Este tipo de conexión es el primer emparejamiento de eventos que se toma para la búsqueda de relaciones temporales. Tales parejas toman como verbo de trabajo al verbo inferior, éste será el que represente a la relación y sobre el cual se exprese la relación temporal; esto quiere decir que para un evento de trabajo **A** y su evento emparejado **B**, la relación asignada siempre tendrá la correspondencia:"*A* **relación** *B*".

El segundo tipo de pareja que se toma en cuenta es la relación de verbos *hermanos*. Consideremos una versión ampliada del ejemplo anterior: *Nos dijo que había llegado el día anterior, vamos a visitarlo*. En este caso se tienen tres *verbos-eventos* dentro de la oración compuesta. En la figura 4-5 se muestra el análisis de dependencias de la nueva oración. En esta ocasión se resaltan los verbos que están en el segundo nivel jerárquico; ambos mantienen una relación de dependencia y temporal con el verbo padre, pero también tienen una relación temporal entre ellos.

Para este segundo tipo de parejas, también se toman en cuenta los *verbos-eventos* de oraciones vecinas. Cambiando un poco el texto anterior, podemos considerar el siguiente fragmento: *Nos dijo que había llegado el día anterior. Haremos planes para visitarlo*. En la figura 4-6 se muestra el análisis de dependencias de las dos oraciones adyacentes. Los verbos superiores se analizan como eventos hermanos y se busca una relación temporal entre ellos. El verbo de trabajo dentro de estas parejas es el verbo de la derecha, es decir, el que aparece segundo en el texto.

En adición a los dos tipos de parejas anteriores, se recurre a una clase de pareja final. Todos los verbos-eventos encontrados también son relacionados con el punto cero, o el punto de enunciación. Dicho punto puede entenderse como el momento de tiempo presente en el cual

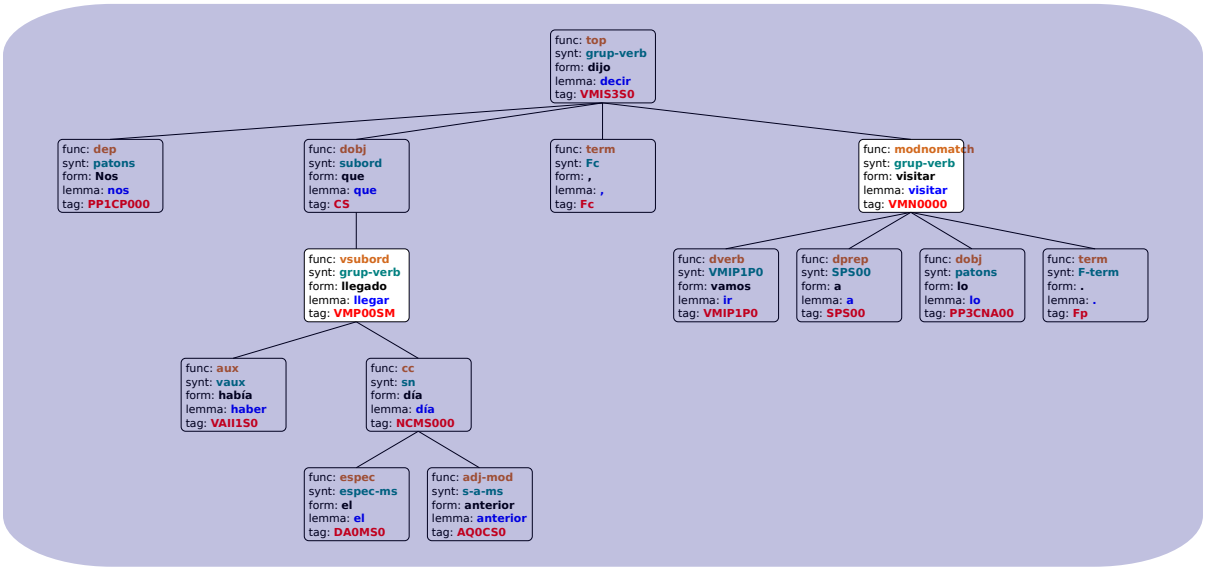


Figura 4-5: Pareja de verbos hermanos

Análisis de dependencias para la oración: *Nos dijo que había llegado el día anterior, vamos a visitarlo.* Se resaltan los nodos del árbol que corresponden a los verbos subordinados de un mismo verbo superior, o bien, a verbos *hermanos*. Además de sus respectivas relaciones con el verbo padre, también se buscará en este estudio una relación temporal entre este nuevo par.

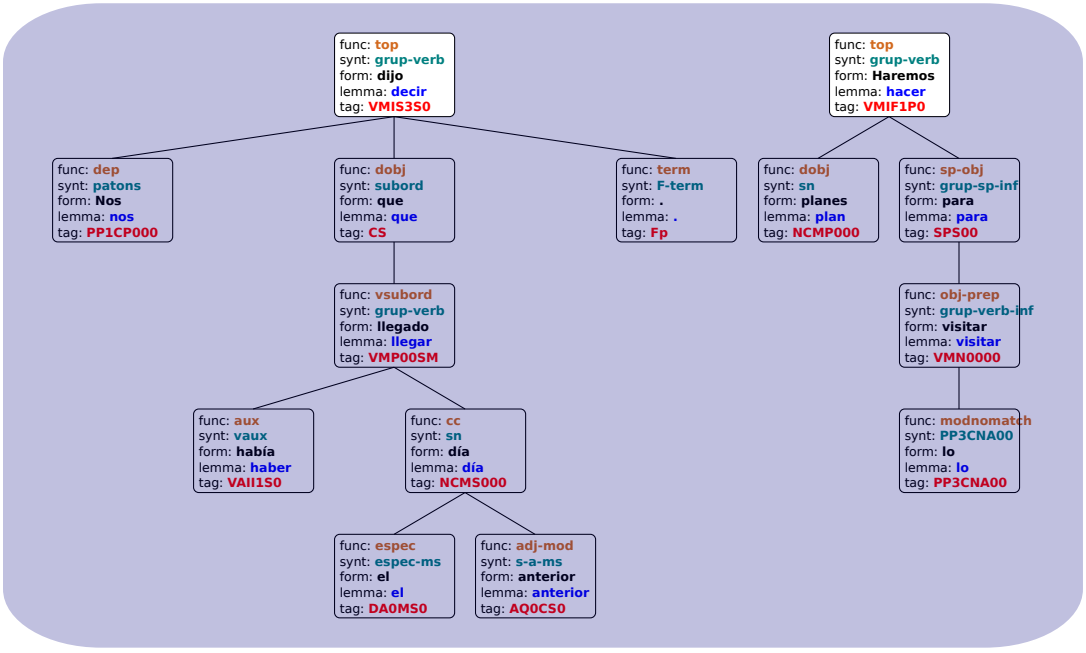


Figura 4-6: Pareja de verbos hermanos en oraciones contiguas

Análisis de dependencia de dos oraciones contiguas. La pareja de verbos-eventos fuera de una misma oración compuesta se logra al tratar a los eventos principales de oraciones adyacentes como si fueran eventos hermanos.

el escritor ubica la narración. En el caso de las noticias, el punto cero suele ser el intervalo de tiempo de la creación del documento.

Las parejas de eventos quedan entonces agrupadas según su conexión. Cada evento tiene al menos una conexión, dada con respecto al origen, y un máximo de tres conexiones al sumarse la pareja con el padre y con el hermano. La conexión total de las relaciones garantiza que no se quede ningún evento no conectado del grafo principal. De esta manera, al seguir un camino de relaciones entre dos eventos cualesquiera con un álgebra temporal como la de Allen [1983], se puede encontrar la relación temporal que mantienen dos eventos aunque no estén directamente relacionados.

4.2.3. Características

Las características que se toman en cuenta para las parejas de eventos dependen de qué tipo de pareja se trata. Muchas de ellas son comunes para todas las parejas en sí, es por ello que en esta sección se mencionan las características que se extraen para las parejas, comenzando por los casos en los que se consideran menos.

Relacion con punto de enunciación

Todos los eventos localizados cuentan con este tipo de relación. Para esto, cada evento no se empareja en sí a otro detectado, por lo que las características extraídas dependen únicamente del verbo que se está analizando. Considerando como ejemplo, la segunda oración de la figura 4-6: *Haremos planes para visitarlo.*

Como se muestra en la figura, el verbo **haremos** es detectado con la etiqueta *'synt:grup-verb'* y colocado en la parte superior del árbol que representa la oración[1]. La etiqueta *'synt:grup-verb'* es empleada para localizar los verbos que son considerados eventos, pero como parte de la recopilación de características, se extraen también las etiquetas *'func'*, *'lemma'* y *'tag'* que acompañan al verbo. En particular, la etiqueta *'tag'* es dividida en cada uno de sus componentes, esto es para tomar en cuenta de forma independiente todos los detalles de la etiqueta *PoS*.

Por último, se extraen los valores de las etiquetas *'func'*, *'synt'*, *'lemma'* y *'tag'* de las

[1]La oración tiene un segundo verbo que se puede detectar como evento: **visitar**. El desarrollo recorre todos los verbos y los procesa uno a uno, en la figura, el verbo no se encuentra resaltado, pero no porque no se tome en cuenta sino porque el verbo de trabajo en el momento es **haremos**.

palabras inmediatamente inferiores al verbo dentro del árbol de dependencias. En el ejemplo, esto correspondería a las etiquetas de las palabras **planes** y **para**.

Para cada palabra, se juntan los valores de sus etiquetas en una sola característica binaria asociada al verbo de interés; es decir que en un vector de características, cuando un verbo tenga subordinada una palabra específica con una combinación de etiquetas específicas, la característica que representa esa combinación tendrá un valor positivo en el vector.

Por lo que respecta al ejemplo, se puede considerar entonces que el vector de características contendrá los valores:

- func_top = 1
- lemma_hacer = 1
- tag_1_V = 1
- tag_2_M = 1
- tag_3_I = 1
- tag_4_F = 1
- tag_5_1 = 1
- tag_6_P = 1
- tag_7_0 = 1
- dobj_sn_plan_NCMP000 = 1
- sp-obj_grup-sp-inf_para_SPS00 = 1

Verbos subordinados

La segunda clase de parejas de verbos, en cuanto a cantidad de características, es la de los verbos subordinados. Este tipo de parejas ya cuenta con dos eventos detectados a partir de los árboles de dependencia.

Al tener dos verbos, se efectúa dos veces la extracción de características utilizada para las relaciones con el punto de enunciación, esto es, extraer las características mencionadas en el punto anterior para cada uno de los eventos detectados. En adición a eso, también se extraen las características de la ruta conectora de los eventos.

Tomando el ejemplo de la figura 4-4 para la oración: *Nos dijo que había llegado el día anterior*. Los verbos **dijo** y **llegado**[2] fueron detectados como los eventos de la oración, para

[2]Se usa **llegado** en lugar de **había llegado** para identificar el evento. El analizador sintáctico reconoce a *llegar* como verbo principal, y al considerarlo nosotros de la misma manera se aprovecha mejor el pre-procesamiento.

cada uno se obtienen las características, de igual manera que para el verbo único relacionado con el punto de enunciación y se agregan al vector de características. La diferencia radica en la ruta conectora entre los dos verbos. Para este ejemplo, el único elemento que se encuentra conectando a los dos verbos subordinados es la palabra **que**.

De manera similar a las palabras subordinadas a los verbos detectados como eventos, se toma el valor combinado de las etiquetas *'func'*, *'synt'*, *'lemma'* y *'tag'* de las palabras que se encuentren conectando ambos verbos y se agregan al vector.

Verbos hermanos

La extracción de características para las parejas de verbos hermanos es igual a la que se hace para los verbos subordinados, es decir, se toman en cuenta todos los datos de los verbos junto con sus elementos inferiores dentro del árbol de dependencias, así como los datos de las palabras de la ruta conectora entre los dos verbos.

4.3. Clasificación

La clasificación de las relaciones temporales se hace en las dos etapas que se muestran en la figura 4-7.

En la primera etapa se hace una separación de los eventos en grupos que corresponden a relaciones temporales similares. Es una categorización gruesa equivalente a manejar menos relaciones temporales. La propuesta se basa en conjuntos de relaciones que aparecieron comúnmente juntas en etiquetados de entrenamiento, dentro de las primeras etapas de planificación y planteamiento de los métodos de relación y clasificación.

Para la segunda etapa, se tiene un clasificador para cada uno de los grupos base de la primera clasificación. Dichos clasificadores solo manejan como salidas las relaciones que forman parte del conjunto. De esta manera, los clasificadores especializados obtienen el resultado de clasificación fina para las trece relaciones finales de Allen [1983].

Entre un par de eventos delimitados siempre existirá una de las relaciones temporales propuesta por Allen. Sin embargo, la complejidad del tratamiento de información temporal en textos radica en la ambigüedad que se presenta tanto al momento de expresar información tem-

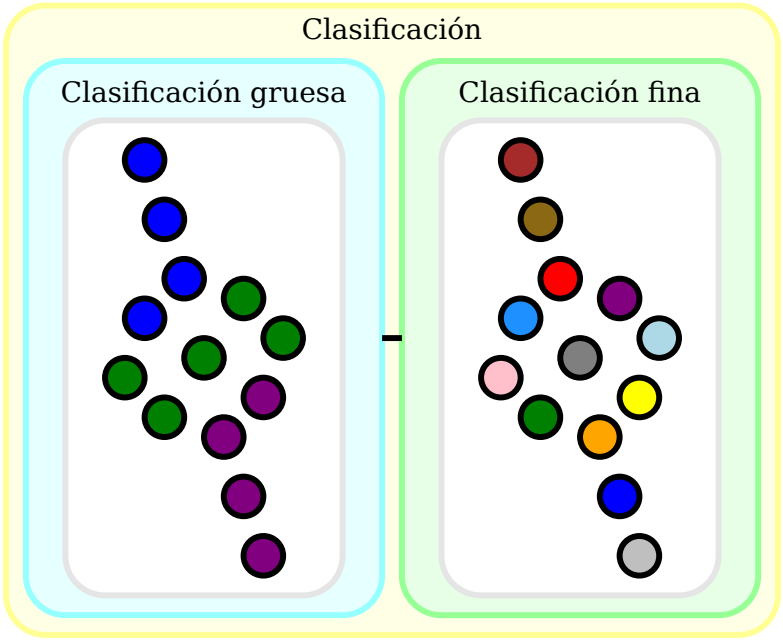


Figura 4-7: Clasificación

La etapa de clasificación consiste en dos pasos, en primer lugar una clasificación gruesa que toma en cuenta un conjunto reducido de salidas y que permite un mejor desempeño para la segunda etapa, que consiste en la clasificación fina de las relaciones temporales.

poral de forma escrita, como al momento de leerla y asimilarla. Dentro de un texto, la mayor parte de la información temporal queda implícita o ausente, abierta a interpretación principalmente cuando el tiempo, de forma estricta, no es el centro del mensaje. En el ejemplo 4.3.1 se muestra una oración en la que han sido localizados dos eventos-verbos. Tal como lo expone Rojo [1990], se puede tener una noción de dirección con respecto a cuál evento ocurrió primero y cuál después, pero la asignación de una relación fina como las de Allen se vuelve muy complicada y polémica.

Ejemplo 4.3.1 Considere la siguiente oración:

- Yo **leía** tranquilamente cuando **sonó** la alarma

En este ejemplo se encuentran dos eventos de forma automática en la oración. Desde el punto de vista de Rojo [1990], la relación se puede categorizar como de actividades simultáneas. Sin embargo, desde el punto de vista de las relaciones de Allen, se pueden argumentar tres relaciones diferentes.

- sonó **durante** leía. (**d**)
- sonó **termina a** leía. (**f**)
- sonó **justo después de** leía. (**mi**)

Las representaciones gráficas de dichas relaciones utilizando los diagramas propuestos por Freksa se muestran en la figura 4-8. No hay información para discriminar la relación que hay entre el extremo derecho del intervalo de los eventos, pues puede ser percibida como que la alarma sonó durante la lectura sin interrumpirla o bien, que el sonido de la alarma interrumpió la lectura, y como última posibilidad, por continuidad, que la lectura y la alarma terminaron al mismo tiempo.

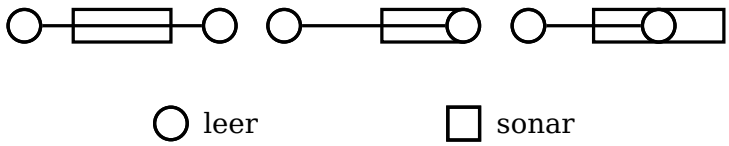


Figura 4-8: Interpretación de variable de relaciones

Posibles interpretaciones gráficas de los eventos *leer* y *sonar* del ejemplo 4.3.1

En otras ocasiones, es común que la misma naturaleza de los eventos impida una correcta identificación de la relación precisa que guardan. En el ejemplo 4.3.2 se muestra un fragmento de oración en el que se han detectado dos eventos.

Ejemplo 4.3.2 Considere el siguiente fragmento:

- ... las canciones de cuna y la música infantil frecuentemente **se basan** en patrones. Muchas de estas canciones también **se hacen** con movimientos como ...

Una vez más, es factible tener una noción de simultaneidad, pero las situaciones que se mantienen a lo largo de un amplio periodo de tiempo, es decir, son estados, se desconoce la relación que puedan tener los límites de sus intervalos. Lo único que se puede saber con certeza es que tienen un espacio en el cuerpo del intervalo de los eventos que es contemporáneo. En la figura 4-9 se muestran las relaciones temporales compatibles al presentarse dos eventos contemporáneos.

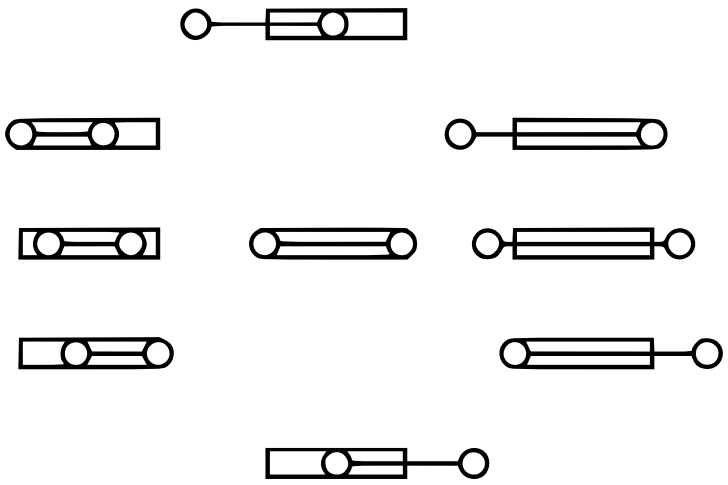


Figura 4-9: Relación temporal entre verbos de estados

Al tener dos eventos del que solo se saben que son simultáneos, se tienen nueve relaciones de Allen compatibles para relacionarlos

La clasificación de las relaciones temporales es, por lo tanto, complicada. El recurso más utilizado es reducir el conjunto de relaciones a categorizar y de esta manera, lograr agrupar las relaciones cercanas o aquellas en las que se tenga incertidumbre. No obstante, dicha práctica representa pérdida de información e incompatibilidad con operaciones como el álgebra de intervalos de Allen [1983], que proporciona aún más información de la que se obtiene en un inicio. En el ejemplo 4.3.3 se muestra una oración en la que se tiene una relación ambigua entre un par de eventos; empero, esta ambigüedad podría desaparecer con el contexto, ya sea en oraciones anteriores o posteriores. El álgebra de intervalos es capaz de desambiguar una relación que también esté conectada con otras relaciones.

Ejemplo 4.3.3 Considere el siguiente fragmento:

- ... decisión judicial y **respaldó** la decisión de la Casa Blanca de **llevar** el caso ante ...

La relación temporal entre *respaldar* y *llevar* es ambigua, puede ser que *llevar* haya ocurrido antes o después del *respaldo*. Sin embargo, es posible que se pueda desambiguar esta situación si surge información extra de otras relaciones que tenga a alguno de los dos eventos involucrados, ya sea *llevar* o *respaldar*.

Con la reducción de relaciones se pierde también precisión de representación cuando se tienen en el texto relaciones sin ambigüedad. En el ejemplo 4.3.4 se muestra un fragmento de oración

en donde se localizaron dos eventos relacionados de una forma clara. Al simplificar el conjunto de relaciones se facilita el etiquetado, pero la categoría en sí, puede tener más interpretaciones que la que el texto proporciona, precisamente porque en una sola categoría se engloban varias relaciones.

Ejemplo 4.3.4 Considere el siguiente fragmento:

- . . . hace cuatro años me **despedí** de OV7, pensando que era definitivo, pero luego **empieza** esta gira . . .

En el fragmento de oración fue detectado un par de eventos. La relación entre los dos eventos es clara, el evento de *despedirse* ocurre antes de que *empieze* la gira, no hay ambigüedad.

4.3.1. Clasificación gruesa

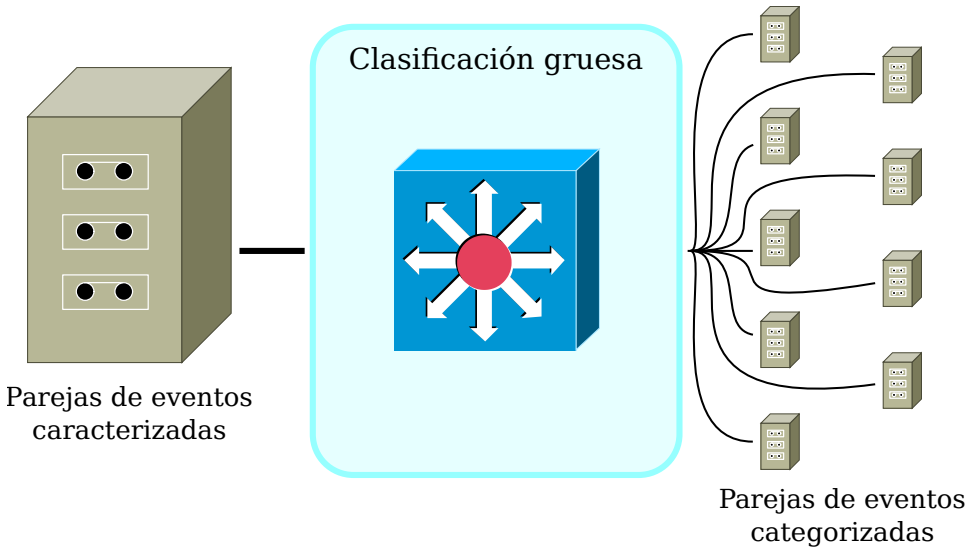


Figura 4-10: Clasificación gruesa

La primera etapa de clasificación consiste en separar las parejas de eventos en grupos que corresponden a relaciones temporales similares para facilitar la clasificación final de la relación temporal entre la pareja de eventos.

En esta tesis se retoma la organización de relaciones de Freksa [1992]. La importancia del vecindario conceptual manejado en dicha representación radica en la forma en que vincula relaciones de manera diferente a las disyunciones de Allen [1983]. Cuando se tiene falta de

información para especificar de forma concreta la relación temporal que existe entre dos eventos, las relaciones posibles son continuas en una vecindad conceptual, con esto se logra evitar la representación independiente de las disyunciones de Allen y simplificar su manejo. La incertidumbre de la relación que se tiene entre dos eventos queda expresada por el conjunto de relaciones adyacentes que delimitan las posibles relaciones que cumplen con la expresión.

Como se ha mencionado, una estrategia usada comúnmente en las investigaciones que manejan relaciones temporales anotadas por personas es trabajar con un conjunto de relaciones reducidas, para lo cual, agrupan relaciones similares y consideran un orden jerárquico estándar en la relación de los eventos, de manera que no es necesario considerar las relaciones temporales inversas de las que se manejan. En contraste, la estructura de relaciones con la que se trabaja en esta tesis, es una aproximación de la estructura temporal a partir de la estructura de dependencia sintáctica del texto. Sin embargo, la jerarquía y el orden de dependencia entre el análisis sintáctico y un análisis temporal son independientes. Esta independencia produce la necesidad de considerar, además de las relaciones temporales básicas, las relaciones inversas o complementarias.

Un caso común que se presenta al momento de reducir relaciones temporales, es cuando dos eventos se traslapan, en donde las dos relaciones: **se traslapa con** (**o**) y **es traslapado por** (**oi**) se representan ambas tan sólo por la relación única de **traslape** [Bramsen *et al.*, 2006a; Chambers *et al.*, 2014; Bramsen *et al.*, 2006b; Kolomiyets *et al.*, 2012]. En el ejemplo 4.3.5 se muestran tres oraciones para explicar el orden jerárquico temporal de los eventos. Con un conjunto reducido de relaciones temporales se corre el riesgo de llevar a cabo una distinción errónea respecto a los eventos señalados entre la primera y la última oración, a menos que se tenga establecido un orden único de dependencia temporal entre uno y otro que indique el orden jerárquico de los eventos dentro de la relación.

Ejemplo 4.3.5 Considere las siguientes oraciones:

- Deje el horno calentando mientras prepara el pastel
- Deje el pastel preparando mientras calienta el horno

Dos oraciones con la misma relación temporal pero los eventos invertidos. Para ambos casos el evento *azul* se traslapa con el evento *rojo*. A pesar de que la segunda oración no tiene sentido,

si un sistema con relaciones reducidas se topa con la primera oración, la relación temporal extraída se puede interpretar como cualquiera de las dos oraciones.

- Mientras prepara el pastel deje el horno calentando

A diferencia de las dos oraciones anteriores, en esta última el evento *azul* es traslapado por el evento *rojo*. La representación debe entonces quedar igual a la de la primera oración, ya sea dándole orden jerárquico diferente, o bien, invirtiendo la relación.

De manera similar a las vecindades de Freksa, en esta tesis se propone una agrupación de las relaciones de Allen para reducir la cantidad de categorías a usar en una primera etapa de clasificación. Como se aprecia en las figuras 3-12 y 3-13, hay vecindades que son muy comunes y es fácil que se presenten, no solo en las salidas del álgebra de relaciones, sino también al momento de interpretar las relaciones temporales dentro de un texto.

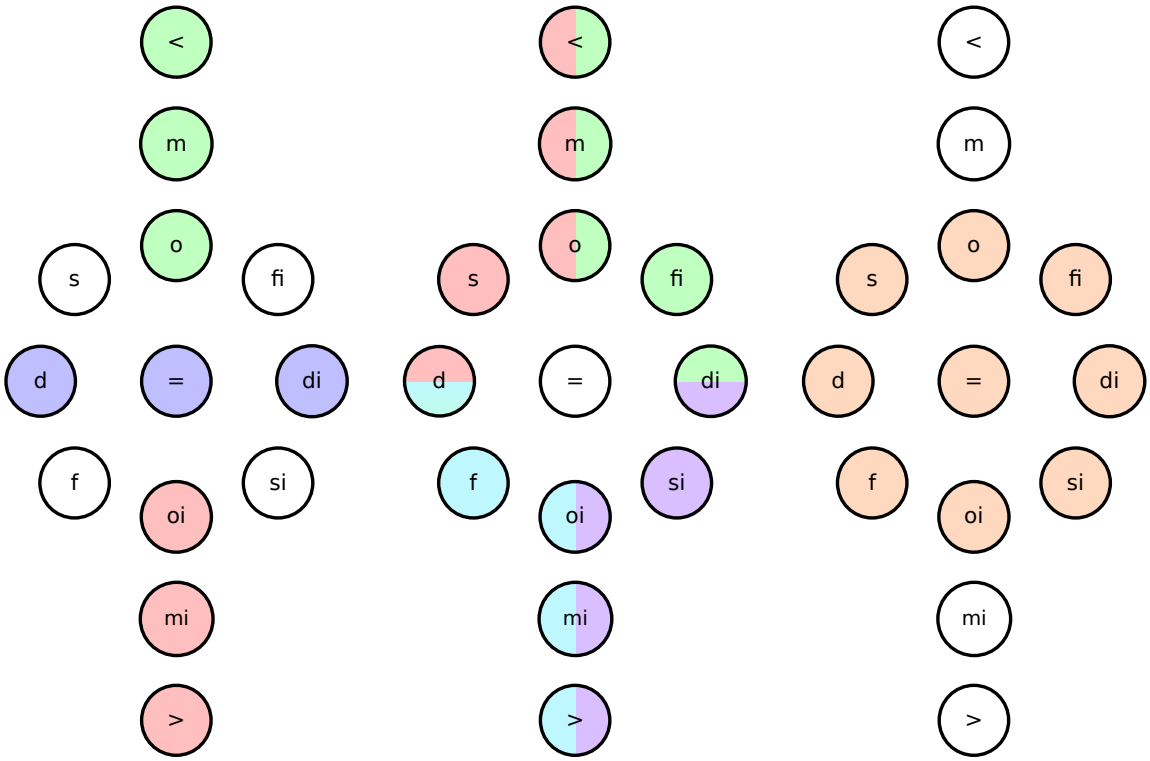
Es importante notar que varias de estas vecindades comunes pueden incluir a otras más pequeñas, lo cual se puede utilizar para agrupar vecindades de relaciones y obtener un conjunto pequeño de vecindades sobre las cuales hacer una clasificación más sencilla. En la figura 4-11 se muestra un diagrama con la propuesta que se hace de vecindarios de relaciones utilizada en esta tesis para manejar una primera clasificación gruesa de las relaciones temporales. Esto con el objetivo de mejorar el desempeño del sistema desarrollado al usar, en primer lugar, un enfoque similar a la reducción de relaciones.

En esta figura, los conjuntos propuestos para la primera etapa de clasificación aparecen representados por colores en la estructura de relaciones propuesta por Freksa [1992]. Se presenta dicha estructura tres veces para que sea más sencilla la visualización de los conjuntos que estamos proponiendo. A continuación se mencionan con más detalle usando los nombres que aparecen en la figura 3-13.

- En la estructura de la izquierda, se muestran tres conjuntos de relaciones.

  "Más viejo y sobrevivido por" que incluye a las relaciones *antes de* (**<**), *justo antes de* (**m**) y *se traslapa con* (**o**).

  "Incluye o inlcuido" que incluye a las relaciones *durante* (**d**), *igual a* (**=**) y *contiene a* (**di**).



(a) Primer nivel de conjuntos
Tres conjuntos de relaciones para determinar las relaciones anteriores, posteriores e incluyentes.

(b) Segundo nivel de conjuntos
Cuatro conjuntos de relaciones para determinar las relaciones traslapadas.

(c) Tercer nivel de conjuntos
Grupo de relaciones de contemporaneidad.

Figura 4-11: Conjuntos propuestos para la etapa de clasificación gruesa

Ocho conjuntos diferentes de relaciones se proponen para agrupar relaciones que aparecen juntas comúnmente según se vio en experimentos preliminares. Las relaciones se categorizan según el conjunto que las incluya totalmente. Los diagramas ordenan los grupos de menor a mayor de izquierda a derecha. Al asignar una etiqueta de relaciones a un conjunto, la búsqueda se realiza en ese mismo orden.

"Más joven y sobrevive a" que incluye las relaciones *es traslapado por* (**oi**), *justo después de* (**mi**) y *después de* (**>**).

- En la estructura central, se muestran cuatro conjuntos de relaciones.

  "Más viejo que" que incluye a las relaciones *antes de* (**<**), *justo antes de* (**m**), *se traslapa con* (**o**), *es terminado por* (**fi**) y *contiene a* (**di**).

  "Más joven que" que incluye a las relaciones *durante* (**d**), *termina a* (**f**), *es traslapado por* (**oi**), *justo después de* (**mi**) y *después de* (**>**).

  "Es sobrevivido por" que incluye las relaciones *antes de* (**<**), *justo antes de* (**m**), *se traslapa con* (**o**), *comienza a* (**s**) y *durante* (**d**).

  "Sobrevive a" que incluye a las relaciones *contiene a* (**di**), *es comenzado por* (**si**), *es traslapado por* (**oi**), *justo después de* (**mi**) y *después de* (**>**).

- En la estructura derecha, se muestra el último conjunto de relaciones, el conjunto **contemporáneo con**, que incluye a las relaciones *se traslapa con* (**o**), *comienza a* (**s**), *es terminado por* (**fi**), *durante* (**d**), *igual a* (**=**), *contiene a* (**di**), *termina a* (**f**), *es comenzado por* (**si**) y *es traslapado por* (**oi**).

La primera categorización de relaciones se hace entre nueve conjuntos de salida, los ocho conjuntos que fueron propuestos en la figura 4-11, y un conjunto extra que comprende a todas las relaciones y que funciona como opción por defecto cuando no se entra en ninguno de los ocho conjuntos anteriores. De manera esquemática, en la figura 4-10 se muestra la clasificación de las parejas de eventos hacia nueve categorías, dicho proceso conforma la etapa de clasificación gruesa.

4.3.2. Clasificación fina

Para manejar la ambigüedad del tiempo en el lenguaje, se estableció el manejo de las vecindades de Freksa. Las vecindades de Freksa son representaciones secuenciadas de las relaciones temporales de Allen [1983]. Dentro de los diagramas, las vecindades siempre tendrán dos extremos, uno *superior* y uno *inferior*. Como ejemplo de esto se puede ver la figura 3-13.

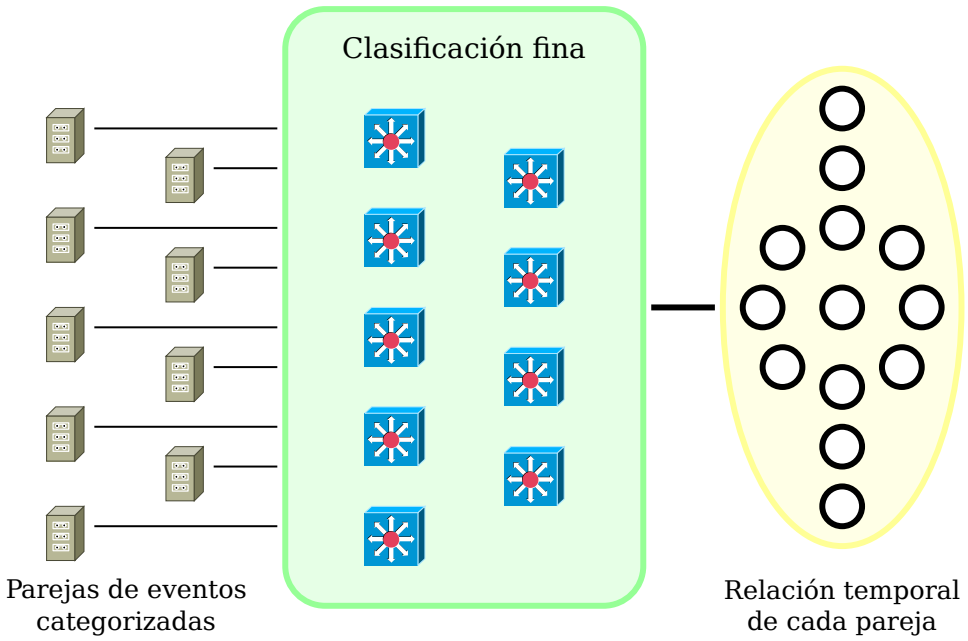


Figura 4-12: Clasificación fina

La segunda etapa de clasificación consiste en utilizar clasificadores especializados en cada uno de los grupos de la clasificación gruesa para encontrar la relación temporal que existe entre cada una de las parejas de eventos.

Aprovechando el hecho de que las vecindades de Freksa son continuas, se plantea la posibilidad de poder identificar las relaciones que las delimitan en sus extremos. Dichas relaciones límite marcan los extremos que engloban la vecindad y, como se observa en la figura 3-13, aportan suficiente información para obtener la vecindad completa.

Dentro de la etapa fina de clasificación, la obtención de las vecindades finales dentro de los grupos base se logra mediante dos clasificadores por cada relación. El primer clasificador se encarga de encontrar la relación límite superior de la vecindad, mientras que el segundo se ocupa de encontrar la relación límite inferior.

Se propone también un algoritmo para el llenado de las vecindades de Freksa a partir de las relaciones límite. Dado que las relaciones límite localizadas pueden no ser compatibles con ninguna de las vecindades encontradas por Freksa, el algoritmo se encarga de generalizar reglas con las que los límites que presenten una relación compatible sean capaces de encontrarla, así como de obtener una vecindad plausible para aquellas relaciones a las que no se las haya agrupado en ningún conjunto.

Completado de vecindades

Todas las relaciones cuentan con dos procesos de clasificación para detectar cada una de sus relaciones límite. La clasificación fina que encuentra la vecindad de relaciones temporales, entre la pareja, se realiza únicamente para dichos extremos, y a partir de esa información se hace un completado de la vecindad para encontrar el conjunto al que pertenece.

Ahora bien, partiendo de las relaciones límite se puede seguir un procedimiento para encontrar el vecindario de relaciones correspondientes a dichos límites. Como se ha mencionado anteriormente, los conjuntos encontrados por Freksa son continuos al considerar las relaciones cercanas como adyacentes. En la figura 4-13 se muestra el diagrama que se toma en cuenta, mostrando conexiones entre los nodos contiguos. Nótese que las conexiones son muy similares a las de la figura 3-11 de las relaciones transformables al cambiar un solo extremo del intervalo. La diferencia radica en que se agregan dos enlaces, uno entre *"durante"* (**d**) e *ïgual a"* =, y otro entre *ïgual a"* (=) y *çontiene a"* (**di**). Estas uniones se dan por la transformación de ampliar o reducir los intervalos en cuestión [Freksa, 1992] y son incluidas para el completado de vecindades debido a la frecuencia de aparición de las vecindades formadas por dichas conexiones.

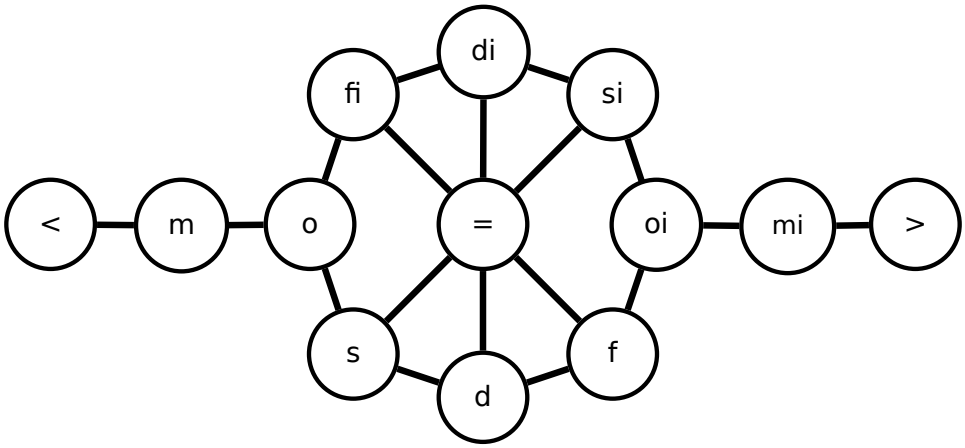


Figura 4-13: Conexiones adyacentes para el completado de vecindades [Freksa, 1992]

En la figura se muestra el diagrama de conexión que se toma en cuenta para hacer el completado de una vecindad a partir de sus relaciones límites [Freksa, 1992].

Las conexiones entre las relaciones funcionan como guías para encontrar las relaciones que comprenden una vecindad a partir de las relaciones límites. El método para completar las relaciones que comprenden la vecindad consiste en obtener todas las relaciones halladas en la o

las rutas que se consideren como la(s) más corta (s) para conectar las relaciones límites.

Con esto resulta factible obtener los conjuntos de Freksa que corresponden en sus extremos con las relaciones límites. Dado el caso de que se encuentre un par de relaciones sin un equivalente en los conjuntos de Freksa, las reglas generalizadas encontrarán una vecindad compatible.

Capítulo 5

Aplicación

En este apartado se expone la implementación de los procesos introducidos en la sección de metodología. Se muestra el proceso seguido para cada una de las etapas de pre-procesamiento, extracción de información y clasificación. Con los procesos una vez concluidos, se presenta también la etapa de resultados y la evaluación del sistema desarrollado.

5.1. Pre-procesamiento

Como se mencionó en la sección 4.1, dentro de esta etapa se trabaja con herramientas lingüísticas automáticas sobre los textos de entrada a fin de conseguir suficiente información para los procesos subsecuentes.

Antes de entrar en detalles sobre la implementación de los procesos de segmentación, etiquetado *PoS* y análisis sintáctico de dependencias, es conveniente describir el conjunto de documentos o corpus que es utilizado en la aplicación de esta tesis.

5.1.1. Corpus

Un corpus es un conjunto estructurado de textos que contienen ejemplos del uso de la lengua que sirven para una investigación [Sierra Martínez, 2015]. En el caso de este estudio, no existía un corpus con la organización y la información necesaria que se pudiera usar como base en un entrenamiento automático.

Como parte de la metodología de esta tesis, se plantea el etiquetado paulatino de un conjunto

de documentos a fin de que funcionen como un corpus de entrenamiento. La adquisición de textos se enfocó en recopilar noticias, debido a la propiedad narrativa de las notas.

Se recopilaron noticias de los portales en red de *Univisión*[1], *Excelsior*[2] y *El Universal*[3] sin ninguna restricción de sección o de tema. Las noticias de *Univisión* y *Excelsior* fueron extraídas el 10 de Noviembre del 2015, mientras que las notas de *El universal* fueron extraídas el 29 de Enero del 2016.

La colección de documentos, es decir, la unión de todas las noticias extraídas de los tres periódicos, consta de 117 notas. En dicha colección fueron detectadas automáticamente, siguiendo la extracción propuesta, 11280 relaciones temporales distribuidas entre 5589 eventos.

5.1.2. Análisis lingüístico automático

Para la etapa de pre-procesamiento se utilizó la herramienta *FreeLing* [Padró, 2012], compuesto por un conjunto de herramientas de código abierto para el análisis del lenguaje. Esta herramienta es capaz de procesar el texto y separarlo en oraciones, segmentarlo, realizar el etiquetado *PoS*, análisis sintáctico de dependencias, y otras tareas de análisis lingüístico.

Los resultados arrojados por *FreeLing* mantienen un formato con la estructura propia del análisis realizado según la consulta. El análisis requerido de *FreeLing*, para este estudio, es el análisis de dependencias; en el ejemplo 5.1.1 se muestra un extracto del resultado de análisis de dependencias realizado por *FreeLing*.

Ejemplo 5.1.1 Un extracto del resultado textual del análisis de dependencias ejecutado por *Freeling*. El fragmento de entrada correspondiente es: "*... la Federación Mexicana de Futbol (FMF) confirmó que el horario del partido que disputarán las selecciones de México y el Salvador* ...".

```
grup-verb/top/(confirmó confirmar VMIS3S0 -) [
  sn/subj/(Federación_Mexicana_de_Futbol federación_mexicana_de_futbol NP00000 -) [
    espec-fs/espec/(La el DA0FS0 -)
  ]
  sn/modnomatch/(FMF fmf NP00000 -) [
```

[1] http://www.univision.com/
[2] http://www.excelsior.com.mx/
[3] http://www.eluniversal.com.mx/

```
  Fpa/modnomatch/(( ( Fpa -)
  Fpt/modnomatch/() ) Fpt -)
]
conj-subord/modnomatch/(que que CS -)
sn/dobj/(horario horario NCMS000 -) [
  espec-ms/espec/(el el DA0MS0 -)
  sp-de/sp-mod/(de de SPS00 -) [
    sn/obj-prep/(partido partido NCMS000 -) [
      espec-ms/espec/(el el DA0MS0 -)
    ]
  ]
  subord-rel/subord-mod/(que que PR0CN000 -) [
    grup-verb/vsubord/(disputarán disputar VMIF3P0 -)
  ]
]
espec-fp/modnomatch/(las el DA0FP0 -)
F-no-c/modnomatch/(    Fz -)
grup-verb/modnomatch/(dará dar VMIF3S0 -) [
    coor-n/subj/(y y CC -) [
    sn/co-n/(selecciones selección NCFP000 -) [
      sp-de/sp-mod/(de de SPS00 -) [
        sn/obj-prep/(México méxico NP00000 -)
      ]
    ]
    sn/co-n/(El_Salvador el_salvador NP00000 -)
  ]
                                        ⋮
```

FreeLing etiqueta cada parte del texto según su *PoS* y sus categorías funcionales dentro del análisis sintáctico.

5.2. Extracción de información

La información de interés que se extrae de los resultados proporcionados por *FreeLing* es la estructura de indentación (que a su vez forma la estructura del árbol de dependencias) y las

categorías asignadas a cada elemento.

Dentro de las etiquetas que maneja *FreeLing* para la identificación de la función sintáctica (primera etiqueta de cada línea en el ejemplo 5.1.1) se tiene un interés particular por localizar los elementos marcados como **grup-verb** y **verb-pass**. *Freeling* asigna una de estas dos etiquetas al verbo que es reconocido como el verbo central de la oración después de realizar el análisis sintáctico; **grup-verb** se utiliza para las oraciones en voz activa y **verb-pass** se utiliza para la voz pasiva. Las palabras claves marcadas con alguna de las dos etiquetas anteriores, así como sus características descritas en la sección 4.2, se extraen como un vector dentro de una matriz para las etapas de entrenamiento y clasificación. Este proceso se lleva a cabo para encontrar todos los eventos del texto, y a su vez, forma la primera "pareja" que se toma en cuenta, pues ésta se hace con respecto al punto de enunciación, por ello se toman en cuenta solamente las características de cada evento individual.

Por último, para completar la etapa de emparejamiento, se realizan dos procesos para interpretar la información estructural de la salida de *FreeLing*.

En el primer proceso, se extrae el árbol de dependencias siguiendo las sangrías de la salida, asignando a cada uno de los elementos sus nodos hijos y sus nodos padres. En la sección 4.2.2 se muestran ejemplos de los primeros árboles mencionados en las figuras 4-4, 4-5 y 4-6

En el segundo proceso, se crea un segundo árbol que contiene información, exclusivamente, de los eventos detectados. Dicho esquema tiene la función de facilitar el proceso de ordenar la información extra para los casos en los que se tienen verbos subordinados y verbos hermanos. En la figura 5-1 se representa un árbol esquemático de eventos con con cinco nodos, es decir, cinco eventos. La raíz representa el texto, por lo que el primer nivel representa los eventos principales de cada oración. Para el caso del ejemplo de la figura se tienen dos oraciones hipotéticas: la primera registra tres eventos mientras que la segunda, dos. Este segundo árbol, sería el resultado de tener la oración de la figura 4-5 y la oración de la figura 4-4 juntas en un mismo texto. En cuyo caso, el árbol de la figura figura 4-5 daría lugar en el árbol de la figura 5-1 a los nodos **A**, **B** y **C**, mientras que el árbol de la figura figura 4-4 daría lugar en el árbol de la figura 5-1 a los nodos **D** y **E**

Una vez que se cuentan con dos árboles de información, el emparejamiento de eventos se logra con un recorrido completo a través de éstos. Navegando por el árbol de eventos, de

izquierda a derecha, y siguiendo por orden alfabético cada uno de los nodos observados en la figura 5-1, es posible encontrar el nodo padre, es decir, el evento al que está subordinado en caso de que no se trate del verbo principal de la oración. Dicha relación se resalta en la figura 5-1 entre los eventos **D** y **E**. La subordinación, es una de las parejas con las que se trabaja, tal como se menciona en la sección 4.2. Las características descritas en dicho apartado, para la pareja ejemplificada, se extraen a manera de un vector dentro de una matriz para las etapas de entrenamiento y clasificación.

A la par del recorrido que se lleva a cabo para emparejar los eventos subordinados es posible encontrar las parejas de verbos hermanos. Después del proceso de búsqueda para el evento padre, se procede a localizar al evento hijo anterior al evento que se está trabajando; de esta manera, se obtiene directamente el verbo hermano del evento de interés. Siguiendo con el ejemplo de la figura 5-1, se muestra este tipo de pareja resaltado entre los eventos **B** y **A**. La relación entre eventos hermanos es también una de las parejas con las que se trabaja; por lo tanto, se extraen las características descritas en la sección 4.2 como un vector dentro de una matriz para las etapas de entrenamiento y clasificación. Con esto se completan las tres clases de parejas que se consideran para los procesos de clasificación.

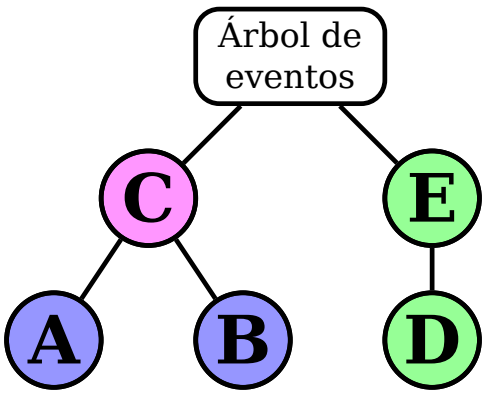


Figura 5-1: Ejemplo de emparejamiento

En la figura se aprecia un ejemplo esquemático de un recorrido ejemplo que se hace sobre el árbol de eventos para hacer el emparejamiento. El recorrido se hace en orden alfabético como se muestra en la figura. Para cada evento, se verifican las tres parejas posibles: subordinación, verbos hermanos, y el punto de enunciación. Esta última se aplica para todos los eventos.

5.3. Clasificación

La última etapa de procesamiento es la etapa de clasificación. La clasificación termina al asignar una vecindad de relaciones temporales para cada pareja de eventos extraídas de un texto dado. Los clasificadores utilizados son entrenados con una técnica de aprendizaje semi-supervisado, lo que quiere decir que el etiquetado y el entrenamiento tienen una relación estrecha dentro del proceso.

En esta sección se explican las etapas de etiquetado y entrenamiento que se utilizan para preparar a los clasificadores, así como la interacción que tienen entre sí. Más adelante, se detalla también la interfaz por línea de comandos del sistema desarrollado. El sistema maneja el proceso de manera iterativa y el entrenamiento puede continuar indefinidamente mientras cuenten con ejemplos para ser etiquetados y el clasificador final conserva todos los datos registrados.

5.3.1. Etiquetado

Como se menciona en la sección 4.2, los pares de eventos se organizan según tres tipos de conexiones (subordinación, hermanos, y punto cero). Cada una de estas clases de relación se maneja por separado al momento de hacer el etiquetado y el entrenamiento. Para cada una de las tres clases se tiene un clasificador que trabaja con las características obtenidas en la etapa de extracción.

El entrenamiento se logra a partir de un etiquetado manual, el cual consiste en el etiquetado fino de la relación temporal de la pareja. Con la información obtenida de las etiquetas finas se obtiene de manera automática la etiqueta gruesa para la relación. Las etiquetas gruesas se utilizan tanto para hacer el entrenamiento del primer clasificador de las relaciones, como para separar las parejas de eventos (junto con sus etiquetas finas) y entrenar por separado a cada uno de los clasificadores especializados.

Para los métodos de clasificación, se establece una numeración que identifica cada una de las trece relaciones de Allen [1983]. En la figura 5-2 se muestra la equivalencia numérica que se utiliza para expresar dentro del sistema desarrollado cada una de las relaciones usadas.

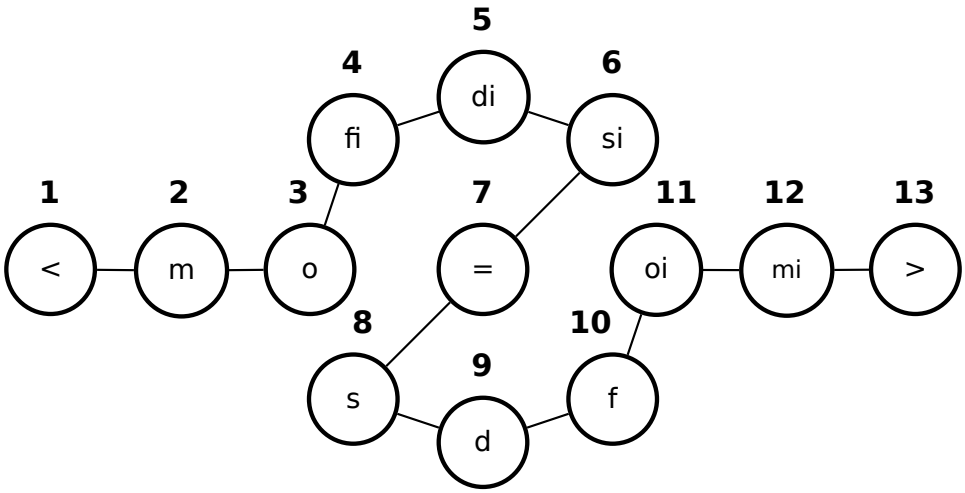


Figura 5-2: Identificadores numéricos para las relaciones temporales

A cada una de las relaciones de Allen [1983] se le asigna un identificador numérico para facilitar su manejo automático.

Etiquetado fino

El proceso de etiquetado se hace con respecto a un fragmento de noticia en el que se haya localizado el par de eventos a etiquetar. Una vez que el sistema ha localizado dicho par, se encarga de obtener un fragmento de texto para presentar en pantalla y recibir la etiqueta correcta. En el ejemplo 5.3.1 se muestra un fragmento de texto con dos eventos localizados. Uno de ellos debe ser detectado como el evento central con respecto al cual se maneje la relación temporal o la vecindad de relaciones temporales.

Ejemplo 5.3.1 Considere el siguiente fragmento:

- ... antes de llegar a su destino el vehículo se metió por un camino de tierra. Ya sobre ese camino los ocupantes se **bajaron** de el taxi e intentaron huir ...

Se detectan de manera automática dos eventos posiblemente relacionados: *metió* y *bajaron*. Con base en la estructura propuesta en la sección 4.2 el verbo *"bajaron"* se detecta como el evento central sobre el cual se define la relación (marcado en negritas a diferencia de *"metió"* que se muestra subrayado). La relación que guarda *"bajaron"* con *"metió"* es de posterioridad, es decir "$>$" en notación de Allen [1983]

Como se introdujo en la sección 4.2, la estructura establece un verbo central para definir la relación entre los eventos. En lo sucesivo, se identificarán en los ejemplos los verbos centrales en negritas, mientras que los verbos secundarios se mostrarán subrayados.

Para los casos que no tienen una sola relación posible, es decir, que su representación corresponde a una vecindad de Freksa, se hace el etiquetado para las relaciones límite. En el ejemplo 5.3.2 se muestra un fragmento de noticia en el que fueron identificados dos eventos. La relación de los eventos dentro del ejemplo no es muy clara, por lo que corresponde a un intervalo de relaciones.

Ejemplo 5.3.2 Considere el siguiente fragmento:

- ... las cuotas partidistas ni los cuates de Los Pinos, tampoco resultan legítimas las campañas de candidatos encubiertos. Y **hay** sospechas de que algunos legisladores que se sumaron ...

La relación entre los eventos detectados corresponde a un intervalo marcado por relaciones contemporáneas. El etiquetado para este par de eventos está dado entonces por las dos etiquetas de las relaciones límite, en el caso de relaciones contemporáneas son las relaciones "**o**" (se traslapa con) y "**oi**"(es traslapado por), en notación de Allen.

El etiquetado de los pares de eventos consiste en las dos etiquetas con las relaciones límite que delimitan la vecindad de relaciones temporales para la pareja de eventos. A partir del etiquetado de las relaciones límite se obtiene de manera automática una tercera etiqueta, la cual contiene la información de la clasificación gruesa.

El proceso que se encarga del etiquetado automático de la clasificación gruesa consiste en dos etapas. En la primera, se completa la vecindad de relaciones a partir de las relaciones límite, mientras que en la segunda se busca a cuál de las ocho categorías gruesas pertenece la vecindad resultado de la primer etapa.

Etiquetado grueso

La clasificación gruesa consiste en asignar en alguna de las ocho categorías la relación de un par de eventos en cuestión. Las categorías gruesas corresponden a una etapa intermedia de clasificación sobre la cual se hace la clasificación fina con un clasificador especializado. Es

por esto que la categoría gruesa depende por completo de las etiquetas que se le asignen a las relaciones límite.

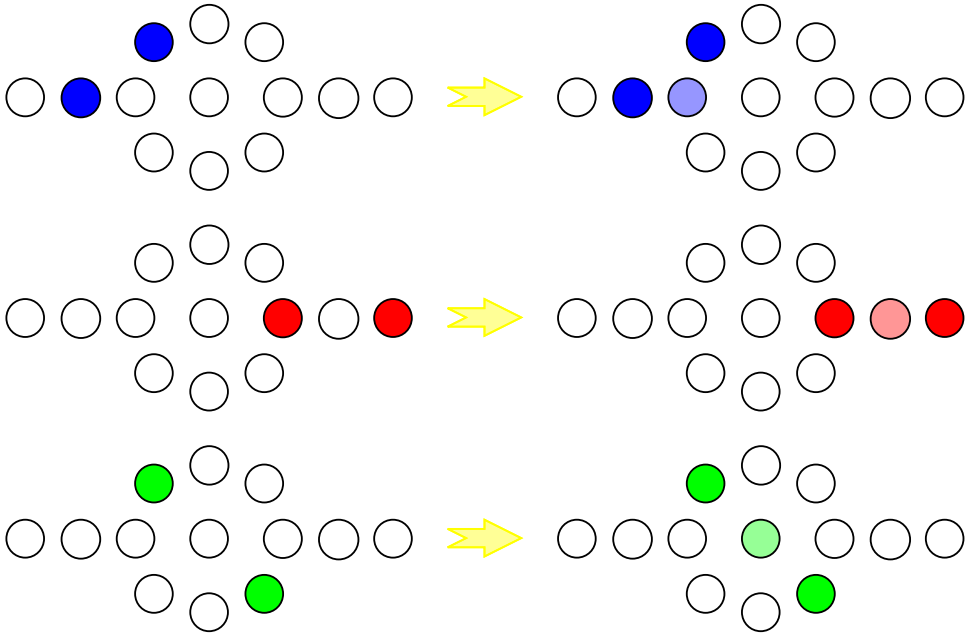


Figura 5-3: Ejemplos de completado de vecindades

En la figura se muestran tres ejemplos en los que una pareja de etiquetas finas son usadas como base para encontrar la vecindad de relaciones temporales posibles entre un par de eventos.

Una vez que se hace el completado de la vecindad, se procede a encontrar cuál de las ocho categorías gruesas es la de menor tamaño pero capaz de incluir a la vecindad completada. En la figura 4-11 se muestran los conjuntos separados en tres grupos: de izquierda a derecha, los conjuntos están ordenados de menor a mayor, es decir, en el diagrama de la izquierda se muestran los tres conjuntos menores que solo constan de tres relaciones; en el diagrama central se muestran cuatro conjuntos que constan de cinco relaciones cada uno; y por último, en el diagrama de la derecha se muestra el conjunto mayor, que consta de nueve relaciones. Los conjuntos son revisados en ese mismo orden para encontrar el conjunto menor que de manera completa incluya un par de etiquetas finas dadas. En la figura 5-3 se muestran tres ejemplos en los que se utiliza una pareja de etiquetas finas, a partir de la cual se completa la vecindad de relaciones para el par de eventos. En la figura 5-4 se presentan nuevamente los tres ejemplos, se muestra la correspondencia entre las vecindades obtenidas a raíz del par de etiquetas finas y

su etiqueta gruesa en cada caso.

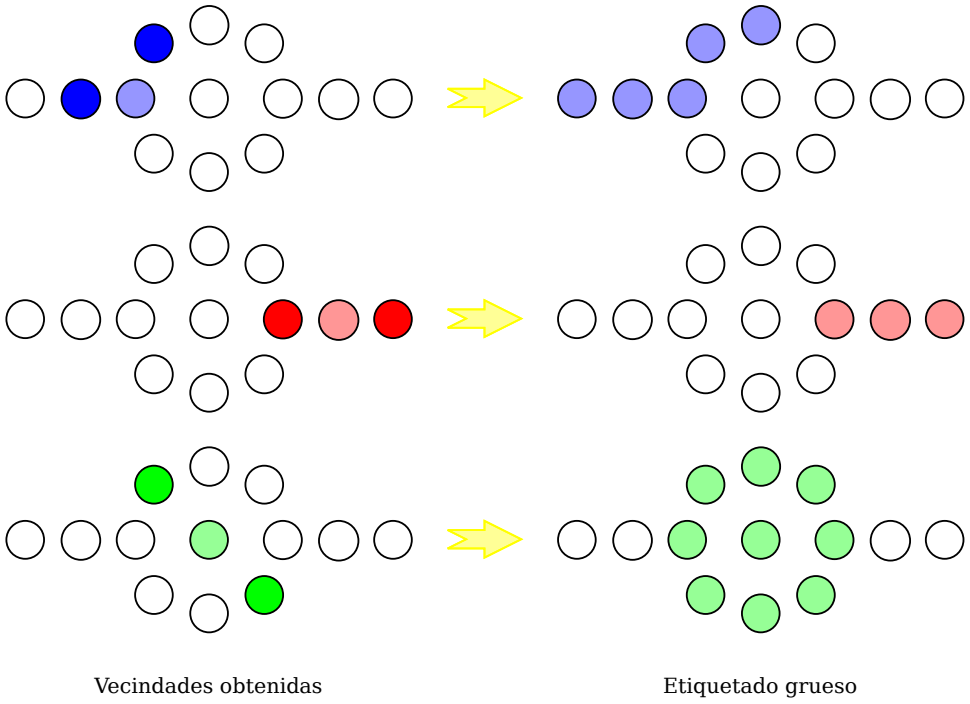


Figura 5-4: Ejemplos de etiquetado grueso

En la figura se muestran tres ejemplos en los que una pareja de etiquetas finas son usadas como base para encontrar el conjunto grueso al cual pertenece.

El etiquetado automático busca y asigna el menor conjunto que contiene a la vecindad armada a partir de las relaciones límite. Para el caso de que ninguno sea capaz de contener por completo la vecindad, se recurre al conjunto de las trece relaciones temporales. Este último conjunto, es el más vago de todos y se usa como un noveno conjunto, en adición a los ocho que ya se habían mencionado.

5.3.2. Entrenamiento

Dado que para este proyecto no se contaba con un corpus etiquetado sobre el que se pudiera hacer un entrenamiento supervisado, a manera alternativa se utiliza el enfoque de aprendizaje semi-supervisado. Como parte de este trabajo, se etiquetó un segmento de un corpus con los grupos de relaciones propuestas y vecindades de Freksa.

De esta manera, las relaciones se etiquetaron con una técnica activa, ésta consistió en la propagación de pocas etiquetas sobre todo el corpus. Como se ha comentado, el aprendizaje

semi-supervisado permite ubicar automáticamente los elementos que tuvieron mayor incertidumbre al ser clasificados. Gracias a esto, se puede establecer un proceso iterativo para hacer anotaciones que aporten la mayor cantidad de información, y reducir así el trabajo de etiquetado; de esta manera se detectan las etiquetas más diferentes a las que se presentan en el etiquetado manual y el sistema las regresa para repetir el proceso de anotación aumentar este etiquetado. El aprendizaje semi-supervisado ha dado los mejores resultados al contar con pocos datos anotados [Zhu, 2005].

El entrenamiento del programa se realizó de forma iterativa junto con el etiquetado. En la primera iteración, se tomaron de forma aleatoria 20 relaciones para ser anotadas y a partir de la segunda iteración se usó la técnica de aprendizaje activo para maximizar el desempeño del sistema hasta lograr una convergencia en la evaluación del sistema.

Con cada relación que el sistema regresa para un nuevo etiquetado manual, se registraron como entradas dos etiquetas, una para cada una de las dos relaciones límite. A partir de las relaciones límite se obtuvo el etiquetado grueso, y todas las etiquetas se agregaron al conjunto de entrenamiento.

Un clasificador se entrena utilizando las etiquetas del etiquetado grueso, y para cada uno de los conjuntos de salida de tal etiquetado, se entrenan dos clasificadores especializados con las etiquetas finas. De dichos dos clasificadores, el primero se entrena con las etiquetas de las relaciones límite superiores y el segundo se entrena con las etiquetas de las relaciones límite inferiores.

5.3.3. Uso del programa

El sistema desarrollado para esta tesis utiliza la numeración de las relaciones para su representación computacional. Cada una de las relaciones puede funcionar como una relación límite que el programa utiliza en toda la etapa de entrenamiento. En la figura 5-5 se muestra la interacción de retroalimentación que tiene el sistema para el proceso de re-etiquetado. El programa muestra una serie de oraciones que considera pertinentes para un etiquetado manual según el algoritmo de entrenamiento semi-supervisado, se muestra en pantalla un fragmento de la noticia de la que se extrajo la relación temporal de interés. Se pueden apreciar dos verbos enfatizados, cada uno de los cuales representa un evento localizado por el sistema. Las etiquetas

esperadas son las que corresponden a las relaciones límite entre el evento resaltado con negritas con respecto al evento subrayado, y las relaciones temporales se muestran según la numeración de la figura 5-2.

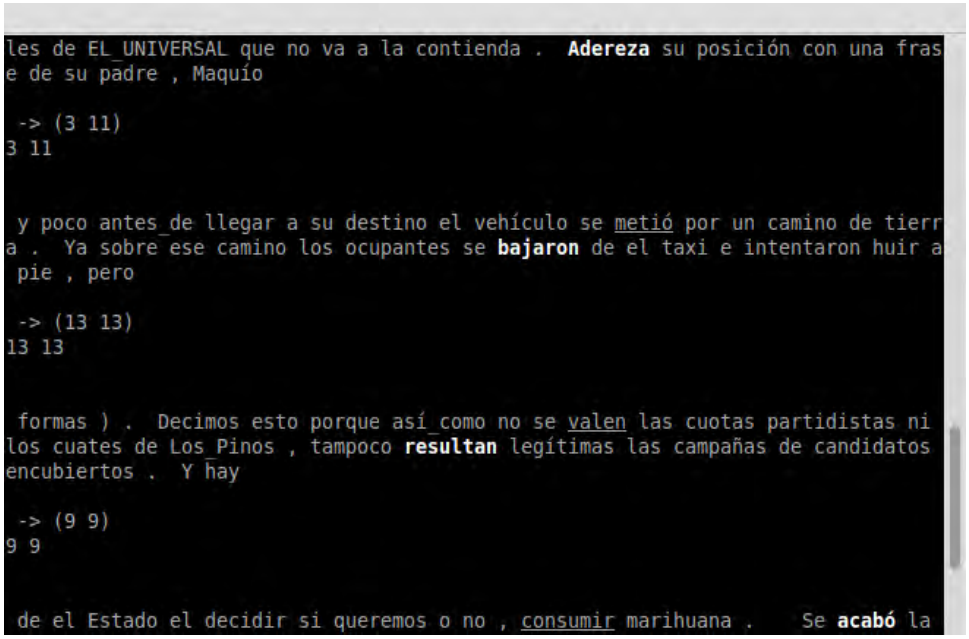


Figura 5-5: Captura de pantalla del sistema

Los principales componentes que se pueden apreciar en cada uno de los fragmentos de noticias son, en primer lugar, un par de eventos resaltados con negritas y subrayados; en segundo lugar, un par de etiquetas, entre paréntesis, que el sistema considera que corresponden a la relación mostrada, y por último se puede observar el par de etiquetas con la que se está re-etiquetando la relación.

En la figura 5-6 se muestran resaltadas las partes que componen la salida del sistema desarrollado. En primer lugar se encuentra el fragmento de la noticia, que toma diez palabras anteriores al primer verbo de interés y diez palabras posteriores al segundo verbo de interés; junto con todo el contenido que se tiene entre los dos verbos se muestra el fragmento para que el anotador interprete, de manera manual, la relación que existe entre ambos elementos. En segundo lugar, se encuentra la salida que el sistema asigna a la relación; el programa siempre asigna una salida para todas las relaciones, siguiendo el sistema de entrenamiento semi-supervisado. Las etiquetas más lejanas a las etiquetadas de forma manual son las que se re-etiquetan. Por último, se tiene la entrada del etiquetador de forma manual: una persona lee el fragmento de la noticia, para entender la relación temporal que existe entre los verbos detectados y establecer las etiquetas que consideren la relación correcta existente entre los dos elementos.

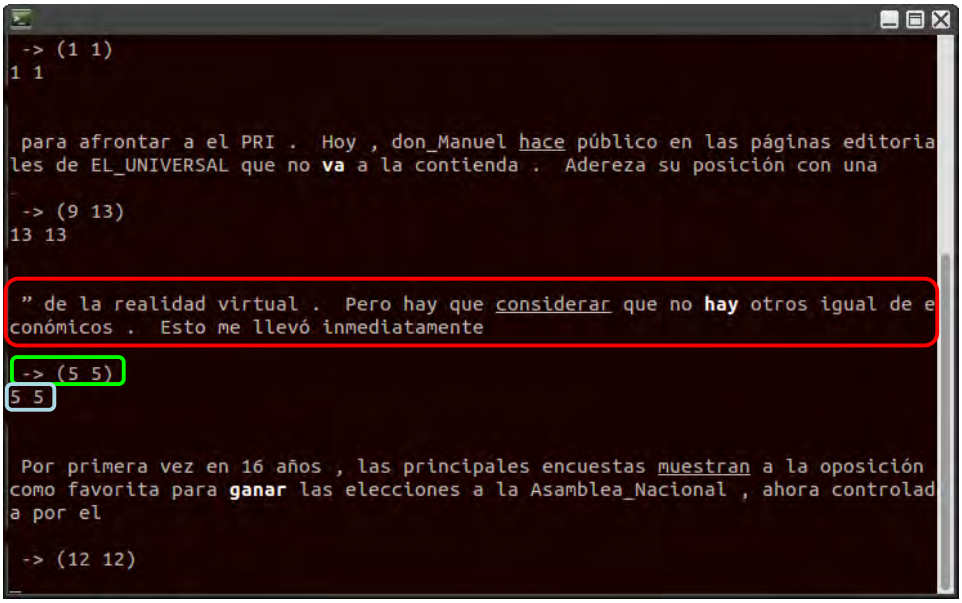


Figura 5-6: Componentes de la interfaz desarrollada

Resaltado en rojo se puede observar el fragmento de noticia que contiene el par de eventos que se está analizando. En verde, se presenta la salida que asigna el programa. En azul, se muestra la entrada que escribe el etiquetador humano, este último dato es una entrada.

En cada repetición del proceso el corpus etiquetado de forma manual aumenta. Los nuevos datos se consideran como verdaderos para hacer la evaluación del desempeño del sistema. Los resultados de la evaluación se obtienen en cada uno de los ciclos del proceso, de esta manera se puede apreciar el progreso del sistema para analizarlo. En la figura 5-7 se muestra una serie de fragmentos con ejemplos concretos de el proceso de etiquetado, todos los fragmentos fueron extraídos de las noticias en la etapa de etiquetado con el programa y las consideraciones antes expuestas. Esto presenta un ejemplo de los datos que recibe el sistema para entrenamiento y, por lo tanto, de lo que se espera como salida del programa. El ordenamiento se da por pares de eventos, las etiquetas representan las relaciones límites del evento en negritas con respecto al subrayado o con respecto al punto de enunciación en los fragmentos en los que no hay verbo subrayado.

Fragmento	E1	E2
el campus exige una investigación de la agresión que **sufrieron** dos jovenes	1	2
ley para evitar inflación, la ley acutal **regula** juegos	1	5
aunque el anuncio no se ha echo, próximamente **anunciará** que	12	12
los vehículos afectados contienen emisiones que el software **escondía**	8	9
al escuchar la historia de la inmigrante que entrevistó ... "aprendí mucho más de lo que **sabía**..."	2	2
y cuando me **refiero** a que	4	6
incluso los bebés aprenden conceptos matemáticos y todo **comienza** con los patrones	8	8
la aplicación es prácticamente igual a la que se **presenta** en iOS	8	10
eso es lo que **hacemos**	7	7
la embarcación en la que **viajaban** 5 personas, llegó en la noche	4	4
veremos qué nos depara el destino	2	2
el accidente se produjo cuando se **aproximaba** a el aeropuerto	2	4
serán las limitaciones que se **impondrán**	8	8
intentaron huir pero los atacantes **abrieron** fuego y los mataron a todos	10	10
no sagen de donde vienen y que también **sirven** para	11	11
las pruebas recientes que le han **realizado** muestran que	3	3
estás listo para ver	9	9
los días que le **restan** parecen ser muy pocos	8	11
todavía no se tiene la zona ya que la empresa **sigue** en pláticas	10	10
se han venido generando experiencias en donde se **puede**	11	11
recibió miles de reflejos de sol que la gente le **dirigía**	3	5
el proyecto tiene su origen en la profesora que lo **coordina**	6	8
la autoridad sabe que no se ha **divorciado**	1	8

Figura 5-7: Tabla de ejemplos etiquetados

En la figura se muestra una tabla con 23 fragmentos de oraciones tomadas de la etapa de etiquetado. Las parejas resaltadas son encontradas automáticamente por el programa. En la primera columna se muestra el fragmento que contiene el par de eventos a ser ordenados, en negritas el verbo sobre el que cae la relación y subrayado el verbo con respecto al cual está la relación. Las columnas E1 y E2 muestran las etiquetas de la relaciones límites inferior y superior respectivamente. En los fragmentos en los que no hay verbo subrayado, la relación se toma con respecto al punto de enunciación.

5.4. Resultados y evaluación

En esta sección se presenta la evaluación de las diferentes salidas que se obtuvieron de las iteraciones del sistema. Se utiliza *validación cruzada* al obtener los resultados de precisión, exhaustividad y medida-F sobre el llenado de la representación de Freksa [1992].

En adición a los parámetros de evaluación para los clasificadores, se muestra también la comparación de los resultados respecto a un enfoque directo de clasificación, es decir, un ordenamiento directo sobre las relaciones finales sin pasar por el etiquetado grueso. En esta sección se presentan gráficas en las que se puede apreciar el progreso del sistema conforme a la cantidad de etiquetas en el corpus, así como su contraste con el desempeño de un sistema sin doble clasificación.

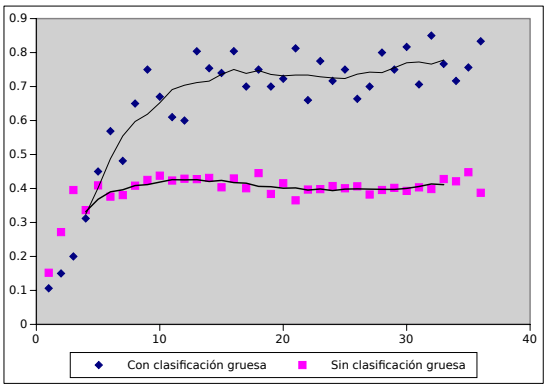
Los resultados obtenidos del entrenamiento y evaluación iterativos se dividen en las tres categorías de relaciones que se introducen en la sección 4.2, es decir, en relaciones subordinadas, verbos hermanos, y relaciones con el punto de enunciación. Como se expone en la sección 4.3, las relaciones son clasificadas en dos etapas, esto es, se ejecuta primero una clasificación gruesa de nueve categorías, y más adelante se utilizan los clasificadores especializados para la categorización final.

Los clasificadores encuentran las relaciones límite del conjunto de relaciones posibles que puede tomar la relación de interés. Los métodos de evaluación se desarrollan sobre las relaciones encontradas una vez que se ha completado el llenado de la vecindad correspondiente. La vecindad completa obtenida por el sistema se compara con la vecindad completa obtenida por el etiquetado humano para obtener los coeficientes de *precisión*,*exhaustividad* y *medida-F*.

5.4.1. Verbos subordinados

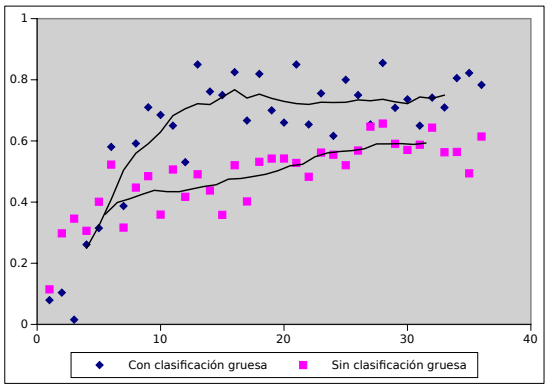
En la figura 5-8 se muestran los resultados de la evaluación para las relaciones de subordinación.

Los valores comienzan a converger después de 15 iteraciones, es decir, con 300 relaciones etiquetadas, y alcanzan una media de medida-F de **72.5 %** al ser clasificados en dos etapas, una clasificación gruesa seguida de una clasificación fina especializada. Cuando se realiza únicamente una clasificación para las salidas finales, sin la clasificación gruesa, se alcanza solo una media



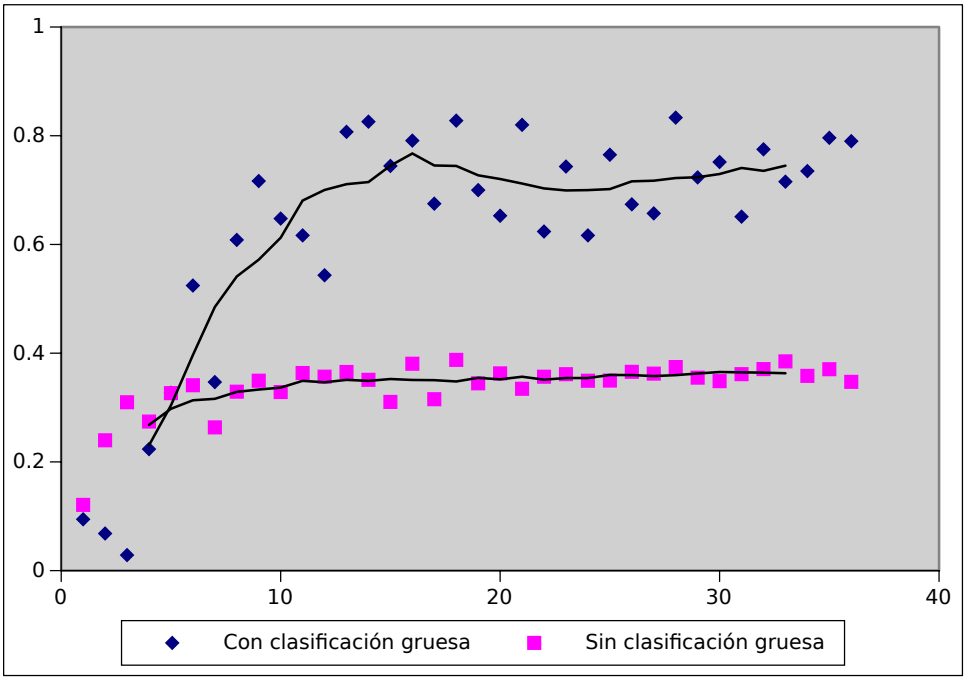
(a) Precisión de relaciones subordinadas

En la gráfica se muestra la precisión que se obtuvo para las relaciones entre verbos que tenían una dependencia sintáctica.



(b) Exhaustividad de relaciones subordinadas

En la gráfica se muestra la exhaustividad que se obtuvo para las relaciones entre verbos que tenían una dependencia sintáctica.



(c) Medida-F de relaciones subordinadas

En la gráfica se muestra la medida-F que se obtuvo para las relaciones entre verbos que tenían una dependencia sintáctica.

Figura 5-8: Evaluación de relaciones subordinadas

Se exponen tres gráficas con los resultados obtenidos para las medidas de evaluación de precisión, exhaustividad y medida-F que se obtuvieron a lo largo del etiquetado y entrenamiento semi-supervisado. En las tres gráficas se muestran en azul los puntos de la evaluación al someter las relaciones primero a la clasificación gruesa de nueve conjuntos y a partir de ello, a una segunda clasificación fina con clasificadores especializados. En contraste, los puntos rosas muestran la evaluación de una clasificación directamente sobre las salidas finales. Las líneas negras muestran las respectivas medias de los puntos de un área cercana.

de medida-F de **36 %**.

En la figura 5-9 se representa el porcentaje de la frecuencia de cada una de las relaciones consideradas para el trabajo. Es importante notar que, para tomar en cuenta toda la información temporal que se está expresando, estos conteos se hacen una vez que el diagrama de relaciones ha sido llenado a partir de las relaciones límite etiquetadas.

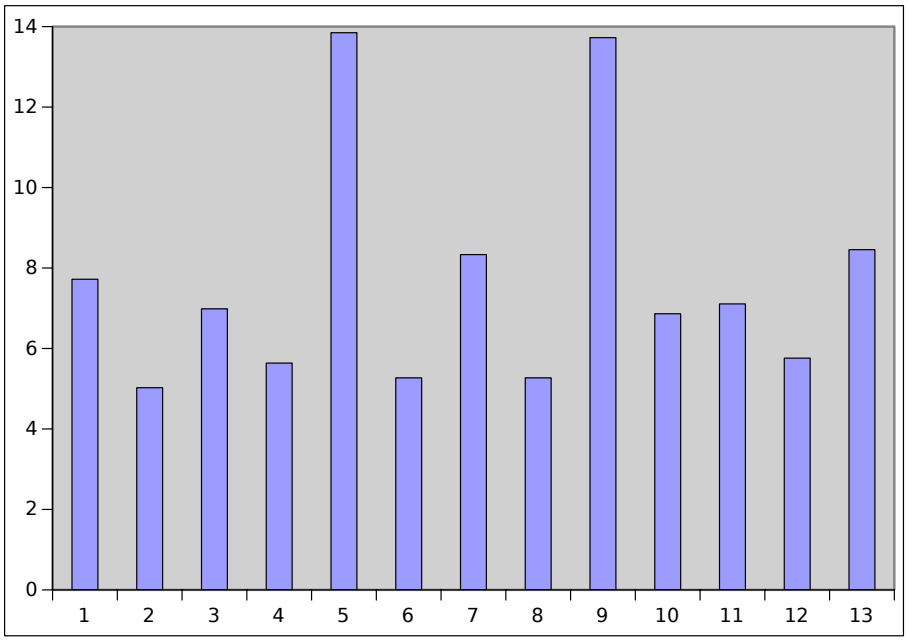


Figura 5-9: Distribución de relaciones subordinadas

Porcentaje de aparición de cada una de las relaciones temporales dentro de las etiquetas. Se puede notar una presencia superior para las etiquetas 5 y 9, que equivalen a las dos relaciones complementarias de *"durante"* (**d**).

Dentro de los conteos de aparición de las relaciones, lo más sobresaliente que se puede observar es la superioridad que tienen dos relaciones en particular. Las relaciones 5 y 9 son las que más aparecen en el caso de los verbos subordinados, y corresponden a las relaciones *ïncluye a"* (**d**) y *"durante"* (**di**) respectivamente. Ambas parecen ser las más utilizadas al relacionar eventos dentro de noticias, es decir, se usan eventos generales de larga duración dentro de los cuales ocurren eventos más puntualizados.

Cuando se tiene el caso en que las relaciones límite que se anotan para una relación son la 1 y 13 (.ªntesz "después de", respectivamente), entonces se puede decir que esa pareja de eventos no está relacionada, o no aporta ninguna información temporal. En el caso de los verbos de

subordinación, tal situación comprende el **4.1 %** de las relaciones etiquetadas.

Del mismo modo, se hizo un conteo de las relaciones que quedaron perfectamente definidas, es decir, las que fueron etiquetadas con una sola relación denotando así una información certera con respecto a la forma en la que estaban relacionados los eventos emparejados. En el caso de los verbos de subordinación, esta característica se presentó en el **68.8 %** de los casos.

5.4.2. Relación con punto de enunciación

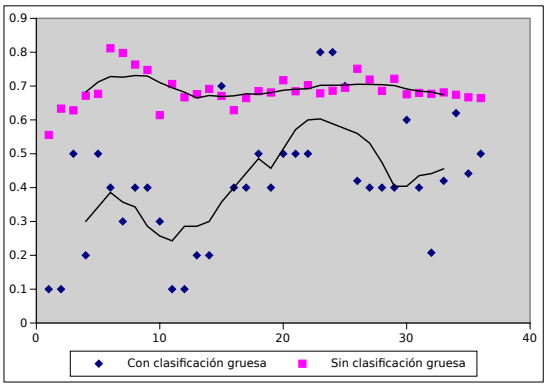
En la figura 5-10 se muestran los resultados de la evaluación para los verbos que fueron detectados como eventos con respecto a un punto de referencia fijo de enunciación, es decir, el momento en el que fue escrita la noticia, el *"presente"* de la nota. Todos los verbos encontrados fueron etiquetados con una relación con respecto a este punto.

Como se puede apreciar en las gráficas, para las relaciones con el punto de enunciación, un paso extra en la categorización no logra obtener resultados positivos. La evaluación muestra salidas azarosas que no convergen hasta el punto del término de etiquetado del corpus. En contraste, las marcas rojas llegan a una salida estable alrededor de la décimo quinta iteración. Para este caso, los resultados del entrenamiento directamente aplicados sobre la salida obtienen valores medios de la medida-F del **67 %**, mientras que la clasificación por etapas varía de manera poco confiable alrededor del **36.5** por ciento.

En la figura 5-11 se evidencia el porcentaje de la frecuencia de cada una de las relaciones que se consideran en el trabajo. Es importante notar que, para tomar en cuenta toda la información temporal que se está expresando, estos conteos se hacen una vez que el diagrama de relaciones ha sido llenado a partir de las relaciones límite etiquetadas.

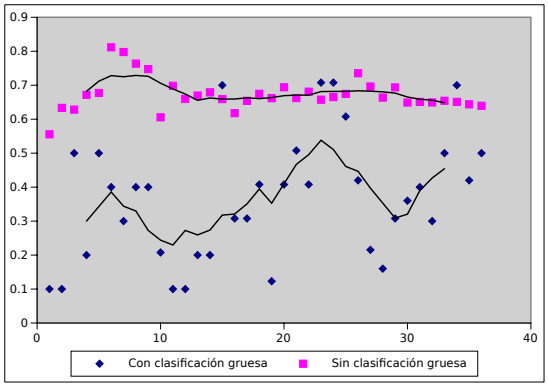
Dentro de los conteos de aparición de las relaciones, se muestran casi exclusivamente sólo tres valores. Las relaciones 1, 5 y 13 son las que más se dan en el caso de las relaciones con el punto de enunciación, dichas relaciones expresan las relaciones *.ªntes"*(<), *ïncuye a"* (**di**) y *"después"* (>) respectivamente. Esto conduce a pensar que las noticias narran, en general, eventos que han pasado, que son una realidad mientras se narran, o que se espera que ocurran. Cuando se relacionan los eventos con el intervalo de tiempo en el que se narra la noticia, las relaciones que se obtienen son equivalentes a las que maneja Rojo [1990].

Con respecto a las relaciones con el punto de enunciación, aquéllas que no están relacionadas



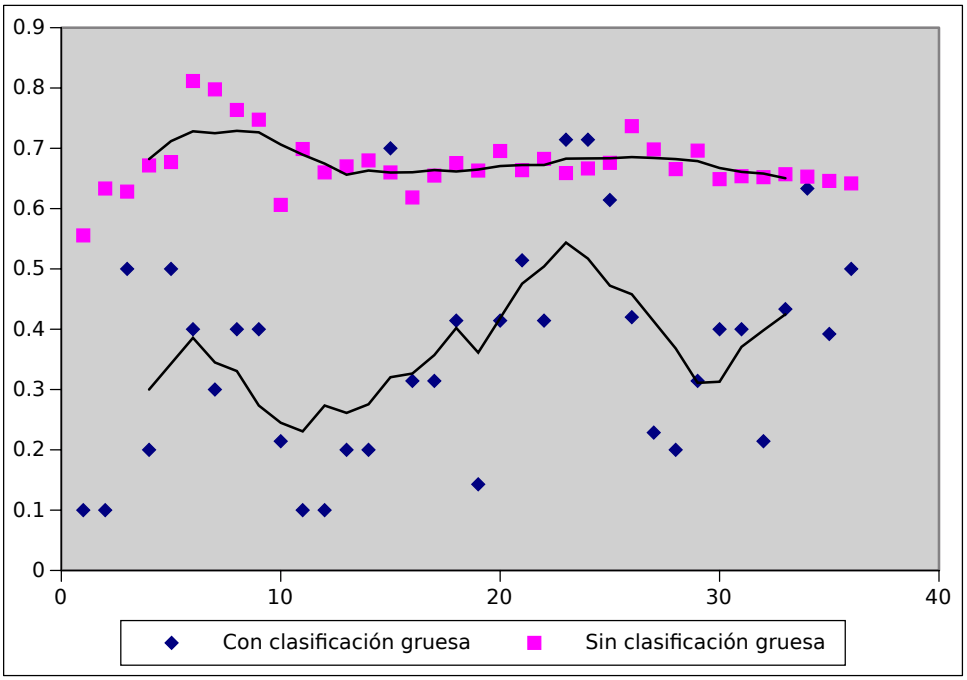
(a) Precisión de relaciones con punto cero

En la gráfica se muestra la precisión que se obtuvo para las relaciones entre verbos y el punto de enunciación.



(b) Exhaustividad de relaciones con punto cero

En la gráfica se muestra la exhaustividad que se obtuvo para las relaciones entre verbos y el punto de enunciación.



(c) Medida-F de relaciones con punto cero

En la gráfica se muestra la medida-F que se obtuvo para las relaciones entre verbos y el punto de enunciación.

Figura 5-10: Evaluación de relaciones con punto cero

Se exponen tres gráficas con los resultados obtenidos para las medidas de evaluación de precisión, exhaustividad y medida-F que se obtuvieron a lo largo del etiquetado y entrenamiento semi-supervisado. En las tres gráficas se muestran en azul los puntos de la evaluación al someter las relaciones primero a la clasificación gruesa de nueve conjuntos y a partir de eso, a una segunda clasificación fina con clasificadores especializados. Los puntos rosas muestran la evaluación de una clasificación directamente sobre las salidas finales. Las líneas negras muestran las respectivas medias de los puntos de un área cercana.

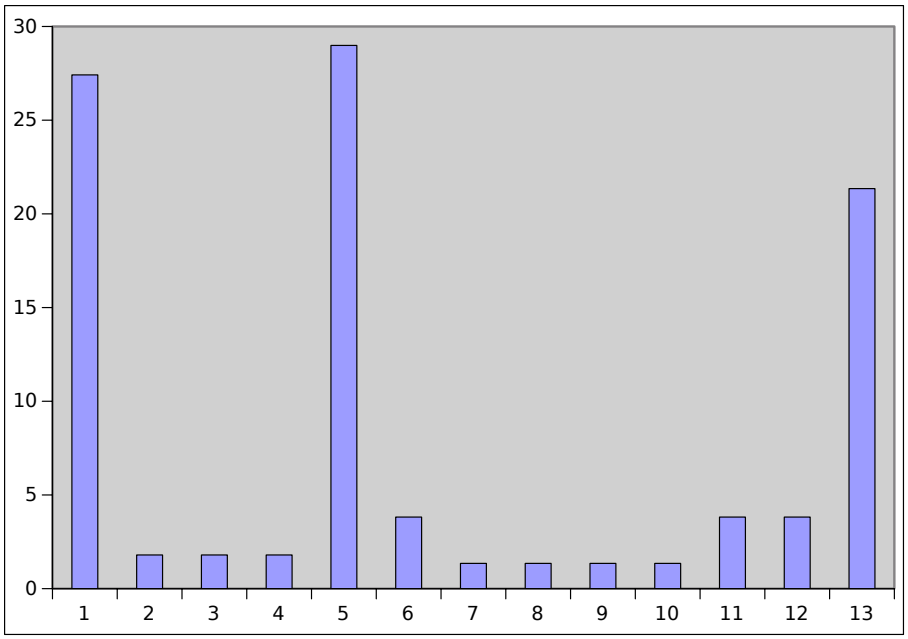


Figura 5-11: Distribución de relaciones con punto cero

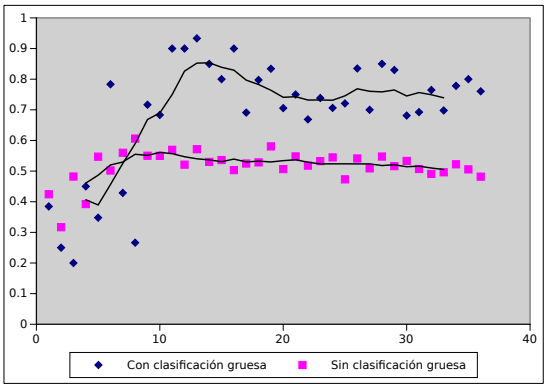
Porcentaje de aparición de cada una de las relaciones temporales dentro de las etiquetas. Se puede notar una presencia casi exclusiva de las relaciones 1, 5 y 13, que corresponden a las relaciones *.ªntes"*(<), *ïncuye a"* (**di**) y *"después"* (>) respectivamente.

temporalmente o no aportan información termporal comprenden el **1.9 %** de las relaciones etiquetadas.

Finalmente, para el caso de los eventos y sus relaciones con el punto de enunciación, también se hizo el conteo de las de las relaciones a los que les fue asignada una sola etiqueta, esto se presentó en el **94 %** de los casos.

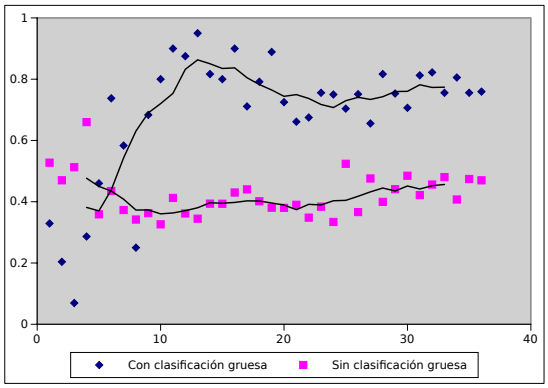
5.4.3. Verbos hermanos

En la figura 5-12 se muestran los resultados de la evaluación para las relaciones entre verbos hermanos, es decir, entre verbos cuyo análisis sintáctico mostró que tenían una raíz común. Para este caso, los valores comienzan a converger después de la vigésima iteración, o sea, con 300 relaciones etiquetadas, y alcanzan una media de medida-F de **72** % al ser clasificados en dos etapas, primero una clasificación gruesa seguida de una clasificación fina especializada. Al realizar una clasificación para las salidas finales, sin la clasificación gruesa, se alcanza únicamente una media de medida-F de **37.5** %



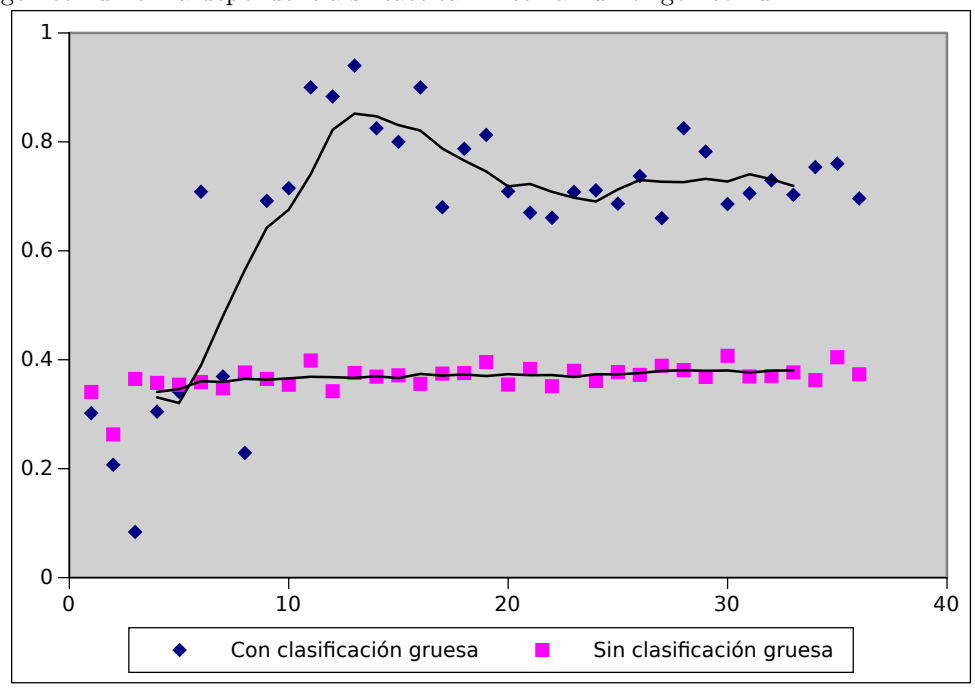
(a) Precisión de relaciones entre verbos hermanos

En la gráfica se muestra la precisión que se obtuvo para las relaciones entre verbos que tenían un origen común en la dependencia sintáctica.



(b) Exhaustividad de relaciones entre verbos hermanos

En la gráfica se muestra la exhaustividad que se obtuvo para las relaciones entre verbos que tenían un origen común.



(c) Medida-F de relaciones entre verbos hermanos

En la gráfica se muestra la medida-F que se obtuvo para las relaciones entre verbos que tenían un origen común en la dependencia sintáctica.

Figura 5-12: Evaluación de relaciones entre verbos hermanos

Se exponen tres gráficas con los resultados obtenidos para las medidas de evaluación de precisión, exhaustividad y medida-F que se obtuvieron a lo largo del etiquetado y entrenamiento semi-supervisado. En las tres gráficas se muestran en azul los puntos de la evaluación al someter las relaciones primero a la clasificación gruesa de nueve conjuntos y a partir de eso, a una segunda clasificación fina con clasificadores especializados. En contraste, los puntos rosas muestran la evaluación de una clasificación directamente sobre las salidas finales. Las líneas negras muestran las respectivas medias de los puntos de un área cercana.

En la figura 5-13 se plasma el porcentaje de la frecuencia de cada una de las relaciones que se consideran para el trabajo. Al igual que los casos anteriores, los conteos se hacen una vez que el diagrama de relaciones ha sido llenado a partir de las relaciones límite etiquetadas.

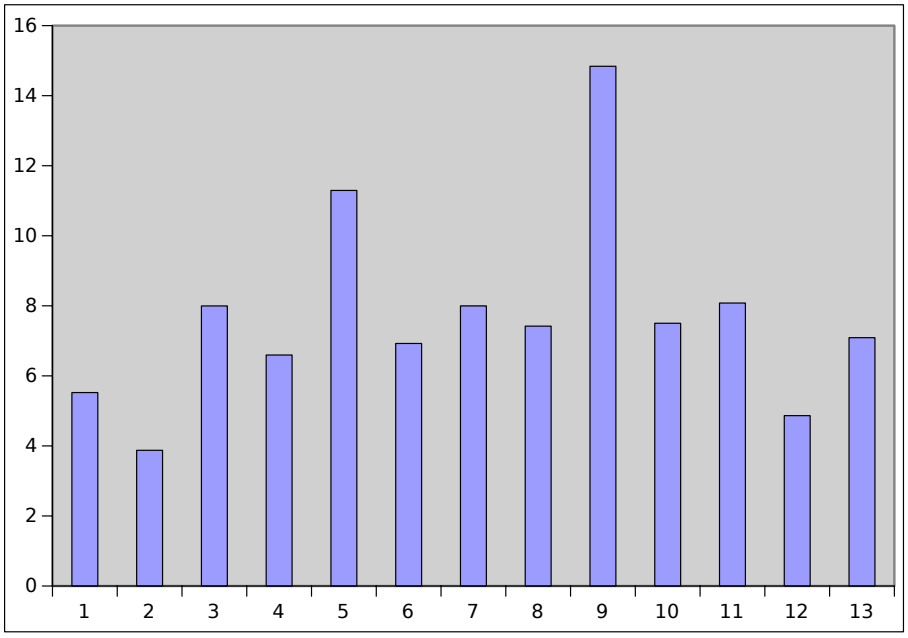


Figura 5-13: Distribución de relaciones entre verbos hermanos

Porcentaje de aparición de cada una de las relaciones temporales dentro de las etiquetas. De forma similar, aunque menor a las relaciones subordinadas, se puede notar una presencia superior para las etiquetas 5 y 9, que equivalen a las dos relaciones complementarias de *"durante"* (**d**).

Las relaciones dentro de esta categoría de parejas verbales es la más uniformemente distribuida; sin embargo, también se tiene una concentración mayor en las relaciones 5 y 9, identificadores que corresponden a las relaciones: *ïncluye a"* (**d**) y *"durante"* (**di**) respectivamente. Estas dos relaciones parecen ser las más utilizadas al relacionar eventos dentro de noticias en diferentes niveles, en otras palabras, se usan eventos para establecer un contexto que puede seguir siendo utilizado incluso en oraciones vecinas.

En cuanto a los verbos hermanos, las relaciones que no están emparentadas temporalmente o no aportan información temporal comprenden el **6.5 %** de las relaciones etiquetadas.

Al igual que en los casos anteriores, se obtuvo la proporción de las relaciones que pudieron ser etiquetadas con una sola categoría. El conteo arrojó una categoría única para el **59 %** de los casos.

Capítulo 6

Conclusiones

A partir de los experimentos anteriores, se pueden obtener las siguientes conclusiones:

Uno de los supuestos que se manejan para esta tesis es que los eventos que están relacionados sintácticamente, también tendrán alguna relación temporal. Dentro de los experimentos, se encontró un máximo de 6.5 % de información temporal nula entre los pares de eventos encontrados; así como un mínimo de 59 % de información temporal completa.

- Las parejas extraídas con relaciones sintácticas detectadas por *Freeling*, también cuentan con una relación temporal.

La clasificación de los verbos al ser relacionados con el punto de enunciación obtuvo mejores resultados al no tener una etapa intermedia de clasificación. Se debe considerar que este tipo de parejas muestran una presencia casi exclusiva de solo tres relaciones, mientras que las relaciones entre verbos presentan gamas más variadas, esto se presenta debido a que las relaciones entre verbo y punto de enunciación son más simples que las relaciones que existen entre verbos.

- Una clasificación intermedia sobre conjuntos de la clasificación final sólo es útil en el caso que se tenga una presencia variada de las segundas categorías, de lo contrario empeora el rendimiento con respecto a sistemas más simples.

En contraste con el punto anterior, los verbos que estaban relacionados con otros verbos mostraron mejora cuando la etapa intermedia fue utilizada.

- El manejo computacional de la estructura en vecindades propuesta por Freksa [1992] mejora el procesamiento automático de la información temporal presente en un texto.

Además de lo anterior, en los resultados se puede apreciar que la distribución de las relaciones que aparecen en los textos no es uniforme, lo cual nos da una idea de la manera en la que se prefieren plantear los eventos que se narran en una noticia.

- La forma más común en la que se relacionan los verbos entre sí es la inclusión de uno en otro.
- Con respecto al momento de la enunciación, los verbos pueden ser categorizados según tres relaciones básicas: antes, después e incluyente. Las otras relaciones temporales se presentan en muy raras ocasiones.

6.1. Trabajo futuro

Dentro de esta tesis se llegan a conclusiones acerca del manejo de las relaciones temporales que pueden servir de base para diferentes trabajos, siguiendo el mismo enfoque de procesamiento de información temporal.

- Este trabajo inicia el etiquetado de un corpus en español con información temporal, y establece pautas para el manejo de dicha información. Sin embargo, carece de una convención o de guías que ayuden a homologar la percepción de los etiquetadores humanos y lograr que el etiquetado sea consistente.
- Se introduce el manejo de una clasificación de relaciones reducidas para emplear una segunda etapa de clasificación para las relaciones finas. Este enfoque podría ser utilizado en conjunto con las relaciones reducidas que se usan comúnmente. De encontrar una equivalencia entre dichas clasificaciones y las vecindades de relaciones temporales, se pueden mejorar los resultados obtenidos mientras se conserva una mayor cantidad de información temporal.
- El enfoque que se toma para la representación de relaciones temporales como intervalos hace incompatible una comparación directa de los resultados del sistema con las evaluaciones del *TempEval*. Un trabajo futuro podría llegar a una transformación para encontrar

una equivalencia entre las representaciones o las evaluaciones y lograr una comparación objetiva.

- Las relaciones más importantes entre verbos son las de inclusión. Con apoyo de un estudio más a fondo sobre la forma en la que se conectan, se puede lograr un mejor desempeño al priorizar la detección y clasificación correcta de este tipo de relación.

- En el caso de los eventos considerados en esta tesis, el aspecto verbal esta directamente relacionado con la amplitud del intervalo temporal que comprende al evento. No obstante, no existe un estudio computacional que permita que se utilice esta característica de manera automática.

- Una de las ventajas de obtener todas las relaciones temporales contra un conjunto reducido, es la posibilidad de utilizar el álgebra de intervalos de Allen [1983] para extrapolar más información. El sistema desarrollado se puede ver enriquecido por la implementación de dicha álgebra sobre cualquier par de eventos.

Bibliografía

Alarcos Llorach, E. *Gramática de la lengua española*, tomo 61. Madrid: Espasa Calpe (1994)

Aldrete, M.C. Gramática de la lengua de señas mexicana. *Estudios de lingüística del español* (28):1 (2009)

Allen, J.F. Maintaining knowledge about temporal intervals. *Communications of the ACM* **26**(11):832–843 (1983)

Allen, J.F. Towards a general theory of action and time. *Artificial intelligence* **23**(2):123–154 (1984)

Allen, J.F. y Ferguson, G. Actions and events in interval temporal logic. *Journal of logic and computation* **4**(5):531–579 (1994)

Allen, J.F. y Koomen, J.A. Planning Using a Temporal World Model. En *IJCAI*, tomo 8, págs. 711–714 (1983)

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L., y Stent, A. Toward conversational human-computer interaction. *AI magazine* **22**(4):27 (2001)

Atserias, J., Comelles, E., y Mayor, A. TXALA un analizador libre de dependencias para el castellano. *Procesamiento del Lenguaje Natural* **35**:455–456 (2005)

Bello, A. *Gramática de la lengua castellana: destinada al uso de los americanos*. A. Roger y F. Chernoviz (1891)

Bethard, S. ClearTK-TimeML: A minimalist approach to TempEval 2013. En *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, tomo 2, págs. 10–14 (2013)

Bolshakov, I.A. y Gelbukh, A. *Computational Linguistics Models, Resources, Applications.* CIENCIA DE LA COMPUTACIÓN (2004)

Botero, L.P. Anterioridad y perfectividad en el sistema verbal en español. *Sintagma: Revista de Lingüística* (9):5–15 (1997)

Bramsen, P., Deshpande, P., Lee, Y.K., y Barzilay, R. Finding temporal order in discharge summaries. En *AMIA*, tomo 2006, pág. 81 (2006a)

Bramsen, P., Deshpande, P., Lee, Y.K., y Barzilay, R. Inducing temporal graphs. En *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, págs. 189–198 (2006b)

Bunger, A. How we learn to talk about events: Linguistic and conceptual constraints on verb learning. *Language Acquisition* **15**(1):69–71 (2008)

Carrera, J., Castellón, I., Lloberes, M., Padró, L., y Tinkova, N. Dependency grammars in freeling. *Procesamiento del lenguaje Natural* (41):21–28 (2008)

Cassidy, T., McDowell, B., Chambers, N., y Bethard, S. An Annotation Framework for Dense Event Ordering. En *ACL (2)*, págs. 501–506 (2014)

Chambers, N. Navytime: Event and time ordering from raw text (2013)

Chambers, N., Cassidy, T., McDowell, B., y Bethard, S. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics* **2**:273–284 (2014)

Delalleau, O., Bengio, Y., y Le Roux, N. Efficient Non-Parametric Function Induction in Semi-Supervised Learning. En *AISTATS*, tomo 27, pág. 100 (2005)

Desclés, J.P. Reasoning and Aspectual-temporal calculus. En *Logic, Thought and Action*, tomo 2, págs. 217–244. Springer (2005)

Dowty, D.R. *Word meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*, tomo 7. Springer Science & Business Media (2012)

Duda, R.O., Hart, P.E., y Stork, D.G. *Pattern classification*. John Wiley & Sons (2012)

Ferro, L., Mani, I., Sundheim, B., y Wilson, G. TIDES Temporal Annotation Guidelines-Version 1.0. 2. *The MITRE Corporation, McLean-VG-USA* (2001)

Filannino, M., Brown, G., y Nenadic, G. ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. *arXiv preprint arXiv:1304.7942* (2013)

Filatova, E. y Hovy, E. Assigning time-stamps to event-clauses. En *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, pág. 13 (2001)

Fodor, J.D. y García, F.A. *Semántica: teorías del significado en la gramática generativa*. Cátedra (1985)

Freksa, C. Temporal reasoning based on semi-intervals. *Artificial intelligence* **54**(1):199–227 (1992)

Galton, A. y Augusto, J.C. Two approaches to event definition. En *DEXA*, tomo 2, págs. 547–556 (2002)

García de Diego, V. Gramática histórica española. *Madrid: Gredos* **395** (1970)

Greene, D., Cunningham, P., y Mayer, R. Unsupervised learning and clustering. En *Machine learning techniques for multimedia*, págs. 51–90. Springer (2008)

Grishman, R. *Computational linguistics: an introduction*. Cambridge University Press (1986)

Group, T.W. *et al.* Guidelines for temporal expression annotation for english for tempeval 2010 (2009)

Guyon, I., Weston, J., Barnhill, S., y Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine learning* **46**(1-3):389–422 (2002)

Hagège, C. y Tannier, X. XRCE-T: XIP temporal module for TempEval campaign. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, págs. 492–495 (2007)

Harris, M.D. *Introduction to natural language processing*. Reston Publishing Co. (1985)

Hermjakob, U. y Mooney, R.J. Learning parse and translation decisions from examples with rich context. En *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, págs. 482–489 (1997)

Jung, H. y Stent, A. ATT1: Temporal annotation using big windows and rich syntactic and semantic features. En *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, tomo 2, págs. 20–24 (2013)

Jurafsky, D. y Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. MIT Press (2007)

Kolomiyets, O., Bethard, S., y Moens, M.F. Extracting narrative timelines as temporal dependency structures. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, págs. 88–97 (2012)

Lloberes, M., Castellón, I., y Padró, L. Spanish FreeLing Dependency Grammar. En *LREC*, tomo 10, págs. 693–699 (2010)

Llorens, H., Saquete, E., y Navarro, B. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, págs. 284–291 (2010)

Llorens, H., Saquete, E., y Navarro-Colorado, B. Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management* **49**(1):179–197 (2013)

Malik, J. y Binford, T.O. Reasoning in Time and Space. En *IJCAI*, tomo 83, págs. 343–345 (1983)

Mani, I. y Wilson, G. Robust temporal processing of news. En *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, págs. 69–76 (2000)

Mani, I., Schiffman, B., y Zhang, J. Inferring temporal ordering of events in news. En *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 HLT-NAACL 2003–short papers-Volume 2*, págs. 55–57 (2003)

Muller, P. y Tannier, X. Annotating and measuring temporal relations in texts. En *Proceedings of the 20th international conference on Computational Linguistics*, pág. 50 (2004)

Padró, L. Analizadores multilingües en freeling. *Linguamática* **3**(2):13–20 (2012)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825–2830 (2011)

Pustejovsky, J. The syntax of event structure. *Cognition* **41**(1):47–81 (1991)

Pustejovsky, J., Saurí, R., Setzer, A., Gaizauskas, R., y Ingria, B. TimeML annotation guidelines. *TERQAS Annotation Working Group* **23** (2002)

Pustejovsky, J., Castano, J.M., Ingria, R., Sauri, R., Gaizauskas, R.J., Setzer, A., Katz, G., y Radev, D.R. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering* **3**:28–34 (2003)

Real Academia Española. *Gramática de la lengua castellana.* Imp. Nacional (1854)

Real Academia Española. *Nueva gramática de la lengua española.* Espasa Calpe Madrid„ Spain (2009)

Reyes Ortiz, J.A. *Metodología para la creación automática de ontologías en Servicios Web: un enfoque basado en perspectivas de la necesidad.* Tesis Doctoral, Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, Morelos (2009). Depto. de Ciencias Computacionales

Reyes Ortiz, J.A. *Creación Automática de Ontologías a partir de Textos.* Tesis Doctoral, Centro Nacional de Investigación y Desarrollo Tecnológico (2013). Dra. Azucena Montes Rendón

Rojo, G. Relaciones entre temporalidad y aspecto en el verbo español. *Tiempo y aspecto en español* págs. 17–43 (1990)

Saquete, E., Munoz, R., y Martínez-Barco, P. Terseo: Temporal expression resolution system applied to event ordering. En *Text, Speech and Dialogue*, págs. 220–228 (2003)

Saquete, E., Munoz, R., y Martínez-Barco, P. Event ordering using TERSEO system. *Data & Knowledge Engineering* **58**(1):70–89 (2006)

Saurí, R., Knippen, R., Verhagen, M., y Pustejovsky, J. Evita: a robust event recognizer for QA systems. En *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, págs. 700–707 (2005)

Schokkenbroek, C. News Stories Structure, Time and Evaluation. *Time & Society* **8**(1):59–98 (1999)

Seeger, M. Learning with labeled and unlabeled data. Informe técnico (2000)

Setzer, A. *Temporal information in newswire articles: an annotation scheme and corpus study.* Tesis Doctoral, University of Sheffield Sheffield, UK (2001)

Setzer, A., Gaizauskas, R., y Hepple, M. The role of inference in the temporal annotation and analysis of text. *Language Resources and Evaluation* **39**(2-3):243–265 (2005)

Shannon, C.E. Prediction and entropy of printed English. *Bell system technical journal* **30**(1):50–64 (1951)

Shannon, C.E. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* **5**(1):3–55 (2001)

Sierra Martínez, G.E. *Introducción a los corpus lingüísticos.* 2015 (2015)

Sowa, J.F. Logical, philosophical, and computational foundations. *Knowledge Representation* (1999)

Strötgen, J. y Gertz, M. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, págs. 321–324 (2010)

Tannier, X., Muller, P. *et al.* Evaluation Metrics for Automatic Temporal Annotation of Texts. En *LREC* (2008)

Torres, X.R. Clases de palabras: El verbo - Modo, aspecto y tiempo. *Lengua Española 1* (2011)

Torres, X.R. Perífrasis verbales. *Español 2* (2012)

UzZaman, N. y Allen, J.F. TRIPS and TRIOS system for TempEval-2: Extracting temporal information from text. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, págs. 276–283 (2010)

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., y Pustejovsky, J. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations (2013)

van Rijsbergen, C.J. *Information Retrieval*. draft edición (1995)

Vargas-Vera, M. y Celjuska, D. Event recognition on news stories and semi-automatic population of an ontology. En *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, págs. 615–618 (2004)

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., y Pustejovsky, J. Semeval-2007 task 15: Tempeval temporal relation identification. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, págs. 75–80 (2007)

Verhagen, M., Sauri, R., Caselli, T., y Pustejovsky, J. SemEval-2010 task 13: TempEval-2. En *Proceedings of the 5th international workshop on semantic evaluation*, págs. 57–62 (2010)

Vicente-Díez, M.T., Schneider, J.M., y Martínez, P. UC3M system: determining the extent, type and value of time expressions in TempEval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, págs. 329–332 (2010)

Vilain, M.B. y Kautz, H.A. Constraint Propagation Algorithms for Temporal Reasoning. En *AAAI*, tomo 86, págs. 377–382 (1986)

Webber, B.L. Tense as discourse anaphor. *Computational Linguistics* **14**(2):61–73 (1988)

Weiss, S.M., Indurkhya, N., Zhang, T., y Damerau, F. *Text mining: predictive methods for analyzing unstructured information*. Springer Science & Business Media (2010)

Wilson, G., Mani, I., Sundheim, B., y Ferro, L. A multilingual approach to annotating and extracting temporal information. En *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, pág. 12 (2001)

Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. En *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, págs. 189–196 (1995)

Yoshikawa, K., Riedel, S., Asahara, M., y Matsumoto, Y. Jointly identifying temporal relations with markov logic. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, págs. 405–413 (2009)

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., y Schölkopf, B. Learning with local and global consistency. *Advances in neural information processing systems* **16**(16):321–328 (2004)

Zhu, X. Semi-supervised learning literature survey (2005)

Zwaan, R.A., Langston, M.C., y Graesser, A.C. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science* págs. 292–297 (1995)