TESIS DE LICENCIATURA

Análisis semántico de Conversaciones Informales

Alumnos:
Isabel Lauria
Luis Eduardo García

Director:
Dr. Alejandro Ariel Vaisman

Diciembre de 2002.

Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales.

Indice

Agradecimientos

Muchos son los que han contribuido a que este trabajo sea una realidad. Es un placer para nosotros hacer un reconocimiento a la ayuda que nos brindaron.
Gracias a nuestras madres, por la atención constante que nos brindaron.
Gracias a Claudio por proveer el hardware que necesitamos, y por tener siempre palabras optimistas.
También fue importante la colaboración de Rosita y Pablo, con la bibliografía y sus consejos oportunos.
Un agradecimiento especial es para Fabián, por incorporar información a la base de datos, por las tardes que nos acompañó con sus mates y por su ayuda incondicional en lo que necesitáramos.
Sería imposible este trabajo sin la presencia del Doctor Alejandro Vaisman, nuestro director de tesis. Gracias a Alejandro por transmitirnos su conocimiento, su experiencia y por guiarnos en la realización de la tesis.
Por último, gracias a Ignacio por su alegría, sus berrinches y con sus 2 añitos fue fuente de inspiración de muchas de nuestras conclusiones.

Capítulo 1

Introducción

En el presente trabajo se analiza el problema de descubrir el tema principal de una conversación dada. Pero no una conversación formal, sino conversaciones en línea, comúnmente conocidas como *chats*.
En este tipo de conversaciones frecuentemente no se respetan gramática, ortografía ni sintaxis. Y en las mismas las ideas muchas veces tienden a ser expresadas de manera extremadamente informal, haciendo más difícil la tarea.
El trabajo estudia cuáles son las características que a pesar de todo predominan en las conversaciones de este tipo. Y a partir de las mismas se desarrolla la teoría sobre la cual se puedan basar herramientas que se dediquen a resolver el problema de orientar y/o identificar el tema principal de una conversación informal. A partir de este análisis teórico, se construye un modelo de herramienta orientada al autoaprendizaje y se estudia su performance en la averiguación de los temas.
El trabajo presenta un enfoque nuevo en el tratamiento de una conversación a partir de patrones de comportamiento de las palabras y frases que se encuentran en la misma.

1.1 Análisis de Comportamiento vs. Análisis Sintáctico

La mayoría de los trabajos existentes en el área intentan determinar un tema a partir de una conversación dada, y a la interpretación del lenguaje formal. Su principal característica es emplear distintas técnicas para obtener una representación abstracta del texto original a nivel sintáctico, y luego compararlo contra información ya almacenada. Que se rijan por reglas de análisis sintáctico, es debido a la presunción de que el texto en estudio respeta la gramática del lenguaje. Otra característica notable es la ausencia de una estructura de relación y/o jerarquías entre los temas almacenados. Tampoco existe un enfoque orientado hacia el autoaprendizaje aprovechando la información existente.
En nuestro enfoque, en cambio, buscamos extender el análisis sintáctico y gramatical tradicional debido a la forma en que normalmente se expresan los usuarios de una conversación en línea. Las personas que participan en la misma tienden a no expresarse correctamente, pero en cambio sí usan palabras que ayudan a los demás usuarios a saber de qué se está hablando. El hecho de que un párrafo u oración esté bien formada o no (bien o mal redactada), o que asevere o no un hecho determinado (que se diga una mentira o

verdad), es independiente del tema de la conversación. Se da mayor importancia a la distribución y presencia de las palabras, aislando de esta manera los problemas asociados a la mala redacción de los párrafos en la conversación.
Partiendo de este enfoque, en sentido teórico, el estudio se centraliza en el cálculo de diversos patrones de comportamiento encontrados en una conversación, y el aporte de los mismos para deducir el tema de la conversación. De tal forma, el objeto en estudio puede definirse como el análisis de la conversación a partir de estos patrones.
La inferencia estadística justifica el acto de decidir cuándo un comportamiento encontrado en una conversación significa información relevante para la determinación del tema. Entonces, se querrá determinar el tema de la conversación cuando se tenga suficiente información para poder hacerlo. Esta determinación se basará en la información aportada por la conversación más la previamente acumulada en una base de conocimiento. En caso de no poseer suficiente información, y según la situación, se le pedirá ayuda al usuario de la herramienta para resolver la situación en particular incorporándose nuevo conocimiento, ampliando la base de conocimiento. La herramienta también inferirá nuevo conocimiento a partir de la nueva información ingresada, ayudando así al autoaprendizaje.
El desarrollo permite que en el futuro puedan analizarse nuevos patrones, e incorporarlos junto a los ya existentes.

Ejemplos 1.1:

- Un centro comercial cuenta con foros de discusión para que sus clientes compartan opiniones. Es de interés del centro infomarse a partir de este medio cuáles son las predilecciones del público para poder adecuar sus ofertas de espectáculos y entretenimientos a esta demanda. ¿La gente prefiere ir al cine o al teatro? ¿Le atraen más los dramas o las comedias? ¿qué obras quisiera volver a ver en cartelera? La herramienta desarrollada, en este caso sirve para poder mantener informado al Centro sobre el ranking de preferencias de sus clientes, y luego poder planificar sus estrategias de marketing con un mayor grado de certeza.
- Un moderador de foros de discusión se encarga de orientar a la gente para que siempre se hable del tema específico del que trata el foro. Y en caso de que esto no suceda tomar una decisión adecuada. La herramienta ayudaría el trabajo del moderador, controlando que siempre el tema de conversación esté relacionado con el tema específico del foro.
- Una empresa tiene a disposición de sus empleados el uso de chats, para que puedan intercambiar sus opiniones en temas relacionados al trabajo que desarrollan. Es de interés de la empresa asegurarse que este medio se utilice correctamente, y que no se convierta en una distracción si es utilizado para otros propósitos, por ejemplo, hablar de fútbol o de política. En este caso, la aplicación sería de utilidad para la empresa, verificando que el chat sea usado para beneficio del trabajo que se esté ejecutando.

1.2 Contribuciones

En este trabajo presentamos una herramienta capaz de analizar una conversación (determinación del tema) sin necesidad de pasar por el análisis sintáctico sino a través del estudio de patrones en las palabras y frases en la conversación.

El diseño está orientado al autoaprendizaje a partir de la información y decisiones del usuario. La aplicación soporta distintos universos temáticos sin restricción. Puede haber sólo temas generales, solamente temas específicos o ambos. Para cada situación habrá una composición distinta de la base de conocimiento que se adapte a la necesidad. Un aporte importante consiste en que la aplicación es capaz de identificar temas a partir solamente de la información asimilada en el autoaprendizaje. En ciertos casos fue necesario adaptar algoritmos basados en datos estadísticos para ser calculados en forma incremental durante la ejecución del parser.
No conocemos otro trabajo que adopte un esquema parecido al desarrollado en esta tesis.

1.3 Marco teórico

El desarrollo de la herramienta se fundamenta en los conceptos de *Redes Semánticas, Razonamiento con Incertidumbre y Teoría de Diseño Orientado a Objetos.*
Las *Redes Semánticas* **[DL90] [JA87] [DK79] [SCH82] [RG82] [RK94]** son un formalismo para la representación de información, y definir las relaciones jerárquicas entre distintos temas. Nuestra aplicación almacena una red semántica en una base de datos apta para la búsqueda y recuperación de los temas tratados, siguiendo la premisa de que un tema es de mayor jerarquía que otro si es más genérico. Cada tema en la red semántica incluye un conjunto de palabras que lo determina de manera única.
Dado un conjunto de palabras destacadas de una conversación obtenida a partir de un parser **[BN00]**, y los distintos conjuntos de palabras asociados a cada tema utilizan técnicas de *Razonamiento con Incertidumbre***[ND97] [RK94]** para la comparación y posterior deducción del tema asociado. Entonces, dados una conversación y un tema se podrá decir con cierto grado de certeza cuándo el tema corresponde a la conversación dada (relación de igualdad), cuándo está relacionado de alguna otra forma con la conversación (relación de padre o hijo) o cuándo no está relacionado con la conversación, usando *Lógica Difusa***[IFS00]**, se puede identificar, con cierto grado de certeza, el tema de la conversación.
El desarrollo de la aplicación se realizó respetando el *Diseño orientado a Objetos***[EG97] [WB90]**, permitiendo la parametrización de los métodos y datos con los cuales el sistema llegará a una decisión. De esta manera el sistema es fácilmente adaptable y modificable.

1.4 Estado del arte

Una gran cantidad de trabajos de investigación de charlas electrónicas o "chats" están enfocados a los problemas relacionados con el mantenimiento y acceso de las estructuras y demás tipos de tareas que surgen del problema de administrar estos foros de conversaciones en internet. En relación a la problemática que nos concierne, existen trabajos de investigación enfocados a la interpretación semántica de textos en lenguaje natural, orientados a distintos dominios, y que emplean distintas técnicas. Todos ellos tienen como premisa principal que los textos originales respeten la gramática del lenguaje, o por lo menos hay una fuerte dependencia con el análisis gramatical.
Sack **[SA00]** desarrolla una aplicación que analiza el texto contenido en foros de Usenet para construir mapas de referencia a estos foros. Los mapas se realizan a partir de la interacción entre los distintos usuarios y analizando el texto relacionado. El trabajo en sí es la suma de

distintas herramientas de las cuales se aprovecha su funcionalidad. En lo que respecta al análisis del texto, y a diferencia de nuestro trabajo, el mismo se realiza a partir del análisis de sentencias dentro de cada párrafo, y la identificación de palabras ligadas entre sí (por ejemplo sustantivos y adjetivos) que se relacionan hilando las respuestas que dan los usuarios entre sí. El análisis semántico se realiza a partir de una herramienta construída para el idioma inglés por Yves Schabes y otros [KSZE88]. Esta herramienta contiene un desarrollo exhaustivo de 317000 formas inferidas más comunes (combinaciones de términos) a partir de 90000 términos básicos.
Otro trabajo del mismo grupo de investigación [SA01], actualmente en desarrollo, trata de solucionar el problema de un usuario al tratar de involucrarse en un mundo de chats. Para esto provee una interface gráfica donde los chats son mapeados de acuerdo a una clasificación previa que se realiza a partir de las relaciones directas entre los participantes de los distintos chats, y un análisis del contenido de las conversaciones.
Al contrario de nuestro trabajo, este análisis de contenido no ocupa un lugar principal dentro del trabajo, y clasifica a las conversaciones de acuerdo a palabras claves distinguidas previamente, pero sin analizar ni informar como se deduce esta información.
Edgar Wendlandt Y James Driscoll [WD00] proponen un método de recuperación de información para la interpretación de consultas de usuarios. El enfoque de este trabajo está basado en conceptos de base de datos de modelos semánticos (particularmente entidades y relaciones entre entidades). Se reconocen propiedades de entidades básicas (atributos) que aparecen en el texto. También se usan conceptos lingüísticos de algo que definen como roles temáticos. Estos roles temáticos permiten reconocer propiedades de relaciones que aparecen en el texto.
Este trabajo no trata el problema desde el punto de vista de hacer un reconocimiento semántico o análisis de contenido de documentos, sino de determinar la existencia o no de determinado contenido en un documento.
Dumais y otros [DFLDH88] desarrollaron una aplicación que está orientado a las consultas que un usuario le formula a una computadora para conseguir información relacionada. Trata de interpretar una consulta (sentido semántico) comparando contra las demás consultas de la base de datos formando vectores ortogonales de acuerdo a una definición dada, y luego descomponiendo en valores singulares.

1.5 Organización de la Tesis

En el capítulo 2 se describe el diseño general de la aplicación, a partir de la definición conceptual de una conversación y el problema planteado, y se define la información necesaria que debe residir en la base de conocimiento y su estructura conceptual, analizándose la variación del contenido de la misma a partir de la incorporación directa de nuevo conocimiento.
En el capítulo 3 se describe teóricamente las necesidades y problemas que debe resolver el parser de la aplicación, y que son determinantes para su diseño.
En el capítulo 4 se define conceptualmente las necesidades que debe cumplir el mecanismo de representación abstracta de la conversación analizada, de manera tal que esta representación pueda ser usada para interpretar la semántica de la conversación. A continuación se plantea el problema de una búsqueda adecuada del tema asociado a la conversación en la base de conocimiento. Se define una métrica para establecer una

relación de orden entre los temas de la base de conocimiento, según su parecido con la conversación analizada.
El capítulo 5 estudia las diversas decisiones que puede tomar la aplicación según el resultado del análisis de la conversación, y las decisiones tomadas por el usuario de la aplicación.
El capítulo 6 describe la arquitectura de la aplicación según la teoría desarrollada.
En el capítulo 7 se detalla la construcción de la herramienta.
El capítulo 8 describe las pruebas masivas realizadas.
Las conclusiones se discuten en el capítulo 9.
El capítulo 10 describe la bibliografía y material de referencia.

Capítulo 2

Representación y Almacenamiento de una Conversación

En este capítulo se plantea el problema y se describe conceptualmente el diseño de la aplicación.

2.1 Planteo del problema

Se necesita determinar la estructura que tiene una conversación a analizar, y a partir de ésta qué información será considerada relevante para deducir el tema asociado. Es necesario entonces definir cómo será el tratamiento de la información y los pasos necesarios para cumplir con el objetivo.

2.2 Solución propuesta

La aplicación leerá la conversación a analizar, rescatando toda información que se considere pueda ser importante para encontrar el tema de la misma. Con esta información se construirá una representación de la conversación original. Las conversaciones en línea pueden ser muy largas, y por ende se necesita que esta representación no sea demasiado extensa. Entonces se construirá una representación abstracta de la conversación analizada. Luego se interpretará esta información resumida para buscar el tema asociado, y analizar los resultados.

Básicamente el diseño de la herramienta tiene 4 componentes principales:

1. Parser predictivo de la conversación.
2. Interpretación de la conversación.
3. Búsqueda de temas relacionados a la conversación.
4. Decisión de resultados.

El *Parser* se encarga de leer la conversación original, filtrar los elementos que no aportan información adicional, y realizar los cálculos necesarios según los patrones elegidos para analizar la conversación.

La *Interpretación* es un análisis mediante el cual se distinguen las características sobresalientes en la conversación, y a partir de éstas, se obtiene una representación abstracta de la misma.

El proceso de *Búsqueda de temas* es el encargado de comparar estratégicamente la representación abstracta de la conversación con los temas almacenados en la base de conocimiento, para encontrar aquellos que sean candidatos a ser uno de los temas buscados.

El proceso de *Decisión* permite elegir entre todos los temas encontrados, cual es el mejor relacionado a la conversación, identificando cual es la relación (si se habla del tema encontrado o de un tema parecido). También se encarga de incorporar el nuevo conocimiento transmitido por el usuario en aquellas situaciones que lo ameriten.

Definición 2.1: [Conversación]
Conversación es un conjunto de párrafos, con un orden secuencial. Cada párrafo está asociado al nombre de un usuario. Muchos párrafos pueden estar asociados a un mismo usuario (ya que cada usuario puede intervenir repetidas veces en la conversación, de manera aleatoria).

2.3 Gráfico del Diseño

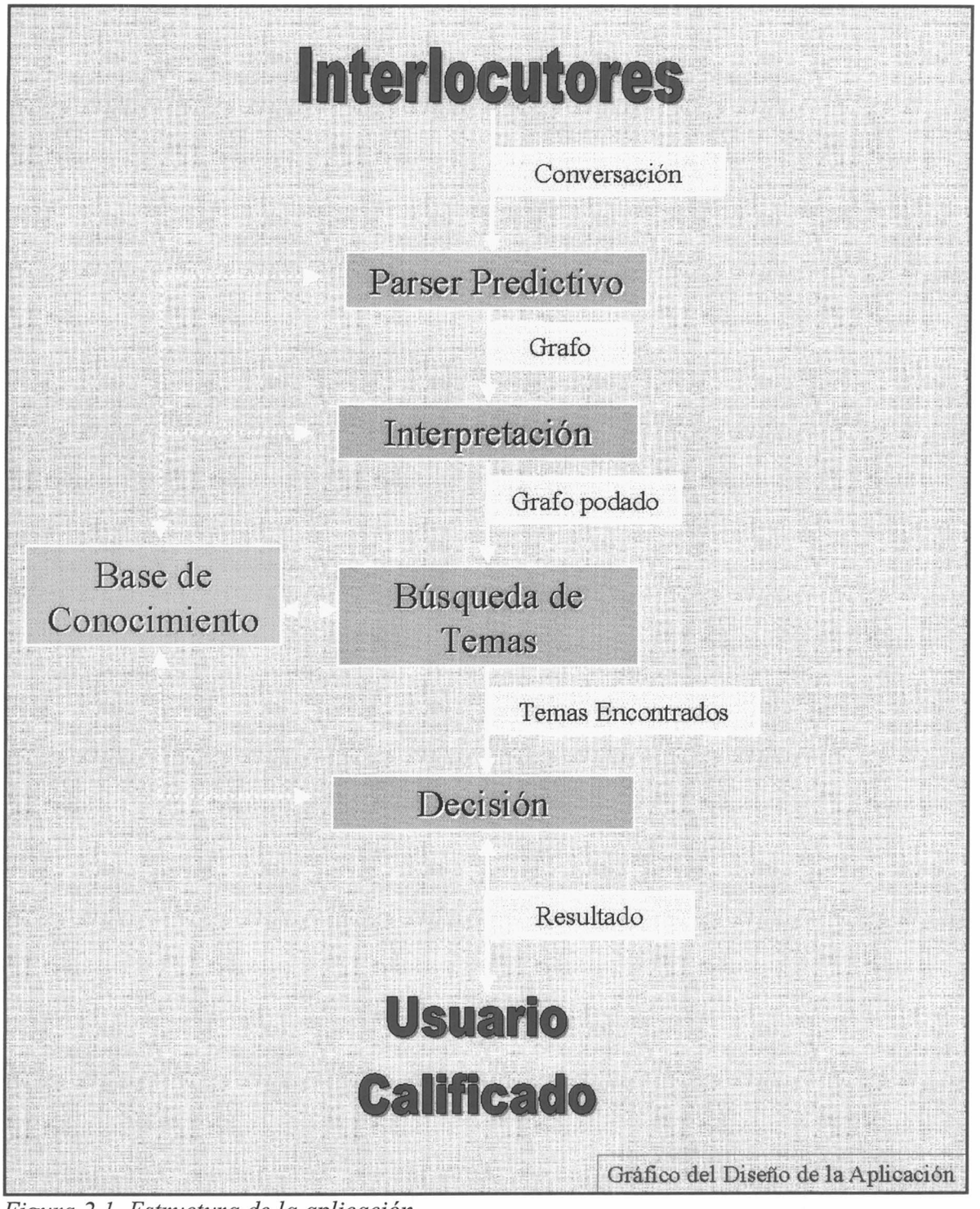


Figura 2.1. Estructura de la aplicación

2.4.1 Red Semántica

La base de conocimiento almacena una *Red Semántica* para organizar los temas de las conversaciones ya que es necesario que los temas conocidos estén contenidos en una estructura que facilite su almacenamiento, recupero y entendimiento. La red semántica está compuesta por nodos, cada uno representante de un tema. Cada nodo tiene la información del nombre del tema, y el conjunto de palabras asociado al mismo. Además, puede haber lazos entre los distintos nodos. Estos lazos son orientados y describen una relación de padre - hijo.

2.4.1.1 Relación tema Padre – tema Hijo

La existencia de un padre implica la existencia de un tema más genérico que el tema relacionado. Esta relación queda determinada a partir de las palabras representantes de un par de conversaciones. Si un tema *A* contiene todas las palabras de un tema *B*, más palabras adicionales, se dice que el tema *B* es *padre* del tema *A*, y el tema *A* es *hijo* del tema *B*.

Ejemplo 2.1:
tema Padre "Música"
Palabras asociadas: "Música", "Canción", "Sonido", "Instrumento"
tema Hijo "Música Rock"
Palabras asociadas: "Música", "Canción", "Sonido", "Instrumento", "Rock", "Guitarra Eléctrica"

2.4.1.2 Temas Distintos

La relación que puede existir entre 2 temas posee restricciones en cuanto a la cantidad de palabras mínimas necesarias para que dos temas sean considerados distintos. Dos temas, identificados a partir del conjunto de palabras que los representan serán considerados *distintos*, si las palabras en que difieren tienen asociada una función (por ejemplo la cantidad de palabras), y este valor supera a una cota mínima establecida para diferenciar cualquier par de temas.

Ejemplo 2.2:
La cantidad de palabras que identifica al tema "Música" es 4, y al tema "Música Rock" es 6. Si la cota mínima para identificar temas distintos es 2, entonces cada conjunto de palabras asociados a los temas "Música" y "Música Rock" podrá identificar al tema correspondiente, sin ambigüedades.

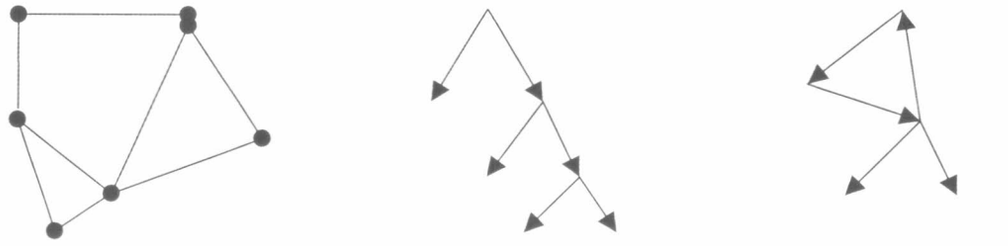
2.4.1.3 Topología de la Red Semántica

Definición 2.8: [Grafo]
Un *grafo* G(V, E) es un conjunto V de nodos, y un conjunto E de pares no ordenados de elementos de V, que se llaman arcos o ejes.

Definición 2.9: [Digrafo]
Un *digrafo* o grafo orientado D(V, E) es un conjunto V de nodos, y un conjunto E de pares ordenados de elementos de V, que se llaman arcos o ejes dirigidos.

Un digrafo tiene un circuito si existen pares ordenados (a_1, a_2), (a_2, a_3), ...,(a_{n-1},a_n), (a_n,a_1)

Definición 2.10: [Jerarquía y digrafo no jerárquico]
Jerarquía es un digrafo sin circuitos.
Digrafo no jerárquico es un digrafo con circuitos



Grafo Jerarquía Digrafo no jerárquico

Figura 2.2. Topologías de redes semánticas.

La topología de la red semántica puede ser:

1. Grafo
2. Jerarquía
3. Digrafo no jerárquico

La elección de la topología de red tiene que considerar la forma en la cual se relacionan los temas almacenados para cumplir el objetivo principal de la herramienta: encontrar el tema más específico del que trata una conversación. La estructura que nos facilita esta búsqueda es una *estructura jerárquica*, que es la que adoptamos en el presente trabajo.
En una jerarquía se puede establecer una relación padre-hijo entre los nodos que forman un arco; el primer elemento del par ordenado es el padre, y el segundo elemento es el hijo.
Un nodo a_i tiene un nivel de jerarquía más alto que un nodo a_j si existen pares ordenados (a_i, a_1), (a_1, a_2), (a_2, a_3), ...,(a_{n-1},a_n), (a_n,a_j).

2.4.2 Variación del Conocimiento

A través del tiempo, existe una relación dinámica de las palabras asociadas a un tema. Las mismas pueden variar dependiendo de las modas y cambios en los usos y costumbres. Hay palabras que dejarían de relacionarse específicamente a un tema, y tal vez serían reemplazadas por otras. O simplemente puede haber una nueva palabra relacionada al tema. Por ejemplo, hace 30 años, relacionado a los peinados se podía mencionar la palabra "gomina". Hoy en cambio, ha sido reemplazada en más de un 99% de los casos por la palabra "gel". Para hablar de los medios de comunicación, se puede escuchar frecuentemente la palabra "mediático", palabra inexistente hace sólo unos pocos años atrás. Estos cambios determinan que la base de conocimiento sea dinámica, no solo en el sentido de incorporar información sobre nuevas palabras y temas, sino también que la misma

El diccionario también se aprovecha para normalizar las palabras que están mal escritas por errores comunes de tipeo, como así también las palabras con faltas de ortografía usuales. Las frases también están almacenadas en el diccionario, porque las frases pueden tener sinónimos y palabras raíces.

Ejemplo 2.5:
"Estados Unidos", "Estados Unidos de Norteamérica", "USA.", "EEUU" son frases que corresponden a la misma clase de equivalencia.

Las palabras o frases que identifican a un tema son Unidades de Información Normalizadas, es decir, son palabras o frases raíces y "representantes" de la clase de equivalencia a la que pertenecen.

Observación 2.1: No se tendrá en cuenta que una misma palabra pueda tener distintos significados (parónimos) porque para ello es necesario hacer un análisis del contexto donde se encuentra la palabra.

Observación 2.2: Algunas palabras pertenecientes a una familia de palabras tienen un significado para el reconocimiento del tema, que difiere del significado de la UIN asociada a esta familia de palabras. La solución que se emplea para que la palabra represente su significado correcto cuando se aplica el método de normalización es asociarle como UIN la palabra o frase que describe su significado, y no la UIN asociada a su familia de palabras.

Ejemplo 2.6:
"corrida" puede referirse a una corrida bancaria. En cambio "correr" o "corriendo" no guardan relación con este concepto. La solución es que la palabra "corrida" tenga asociada la UIN "corrida bancaria" y no "correr".

2.5 Resumen

En este capítulo se presentaron en forma resumida las 4 componentes principales de la aplicación: *parser*, *interpretación*, *búsqueda de tema* y *decisión*, las necesidades que debe satisfacer cada una y las estructuras principales de información que conforman la base de conocimiento: temas y diccionario.
La división del diseño en etapas es fundamental para poder distinguir y encapsular las distintas funcionalidades que interactúan para la resolución del problema. Este encapsulamiento además permite una adaptación a cambios minimizando el impacto de los mismos.
La base de conocimiento debe contener la información necesaria a ser utilizada por la aplicación de forma tal que la misma sea suficiente para cumplir con la demanda de conocimiento. Además debe estar adaptada a cambios inducidos por distintos factores, y debe estar soportada por estructuras de almacenamiento que permitan mantener estos cambios a través del tiempo y facilitar la recuperación puntual de la información.
En el siguiente capítulo se trata la primera componente de la aplicación: el parser predictivo.

Capítulo 3

Parser Predictivo

La función que cumple el parser es recorrer la conversación, analizando la incorporación o descarte de palabras y frases encontradas en la estructura del análisis, filtrando todos los signos de puntuación y caracteres que no aporten información adicional a la conversación para la distinción del tema. En el recorrido, el parser también calcula atributos encontrados en las palabras y frases que servirán para el análisis posterior. Como resultado se obtiene una conversación filtrada.

3.1 Atributos de las palabras y frases

El parser distinguirá dos atributos relacionados al aporte de información que tienen las palabras y frases de la conversación. Estos atributos son:

- *Clave*. Una palabra o frase será *Clave* de algún tema si es muy probable que se hable de ese tema cada vez que esté presente esa palabra o frase. Y a su vez, es muy poco probable encontrar esta palabra o frase en otro tema. Podemos ampliar esta definición para un conjunto de temas en lugar de un solo tema. Así, una palabra o frase será clave de un conjunto de temas si es muy probable que se hable de cualquiera de los temas del conjunto cada vez que esté presente esa palabra o frase. Y a su vez, es muy poco probable encontrar esta palabra o frase en temas fuera de este conjunto. Nótese que no es necesario que los temas pertenecientes al mismo conjunto tengan algún tipo de relación entre sí, pero es de esperarse si la palabra o frase guarda el mismo significado semántico en todos estos temas.

  Ejemplo 3.1:
  Si se menciona "Batistuta" en una conversación, es muy probable que el tema del que se esté hablando sea "Fútbol". También es muy raro que en una conversación no relacionada con "Fútbol" esté presente la palabra "Batistuta".
  Así, la palabra "Batistuta" es palabra clave del tema "Fútbol".

- *Palabra o frase descartada*. Al contrario de las claves, una palabra o frase en condición de ser filtrada será aquella cuya presencia no aporte información adicional para la deducción del tema de cualquier conversación analizada. A partir de esta definición, es

de esperarse que palabras como los artículos, preposiciones, adverbios, verbos y adjetivos comunes sean filtradas de modo tal de no ser tenidas en cuenta en el análisis de la aplicación. Notar que la característica de "para filtrar" de una palabra o frase es independiente de la conversación que se esté analizando.

3.2 Incidencia de las claves

El peso de una clave en un tema tiene relación con la probabilidad de que se hable de este tema cada vez que se menciona esta palabra o frase considerada como clave.

Observación 4.1: Veamos las claves desde otro punto de vista.
Si una palabra es clave entonces se cumple:

Si existe clave en la conversación
$\Rightarrow$ *se habla del tema(i)* *para algún i tq 1 <= i <= n*
(donde n es la cantidad de temas que existen en la base de conocimiento)

$\uparrow$

(se determina con un grado de certeza z que...)

Esto no quiere decir que una clave tenga que estar en todas las conversaciones de un *tema i*. Pero si dicha clave está presente, hay una alta probabilidad de que se hable de ese tema.
Una palabra común (que no es clave) aporta conocimiento como tal, teniendo una valuación especial debido a cada uno de sus patrones, su frecuencia, concentración, distancia promedio, etc. (ver Sección 4.2). Una clave aporta conocimiento como identificadora del tema, más un conocimiento especial que representamos mediante el peso. Además guarda un significado semántico especial en la conversación, por su simple relación con el tema, independientemente de la valuación especial dada por la conversación.

3.3 Parser de una sola pasada

A partir de la existencia de las frases, el parser no considerará palabras hasta no estar completamente seguro que las mismas no forman parte de alguna frase que pueda encontrarse en la conversación analizada. Es decir, una palabra será analizada si no forma parte de una frase que se encuentra en la conversación. Para que esto se cumpla, se necesita tener un parser predictivo. Esto es, debe conocer información "de lo que vendrá" para poder tomar una decisión sobre la situación actual.
El parser es de una sola pasada, y la misma se aprovechará para calcular información adicional necesaria para las siguientes etapas de la aplicación. Esta información está compuesta principalmente por el cálculo derivado de los *Patrones de Comportamiento* en la conversación, que se verán en detalle en la Sección 4.2. Con esto se consigue un ahorro significativo del tiempo de procesamiento de la conversación.

3.4 Conversación Filtrada

Como resultado del parsing, se obtiene un resumen de la conversación original, con la información necesaria y suficiente para realizar el análisis necesario en la determinación del tema. A este resumen lo llamamos *Conversación Filtrada*.

3.5 Resumen

En el presente capítulo se describió la función del parser de la aplicación utilizado para la lectura y análisis primario de una conversación.
El parser predictivo, correspondiente a la teoría de compiladores, permite que el análisis de la conversación se logre en la lectura de una única pasada, ahorrando los costos asociados de procesamiento.
En el siguiente capítulo se tratan la segunda y tercera componentes de la aplicación, que se ejecutan a partir de la información generada por el proceso del parser predictivo: la interpretación y descubrimiento de temas de la conversación.

Capítulo 4

Interpretación y Descubrimiento de Temas

Para analizar la conversación en forma automática, es necesario construir una representación abstracta de ésta. Se requiere que la información que surja a partir de esta representación no sea redundante, y que determine unívocamente a un tema, cuando sea posible, u oriente de manera efectiva hacia la deducción del mismo.

4.1 Interpretación

Una interpretación es la combinación entre una representación abstracta y el contexto de conocimiento (entorno) en el que se analiza la conversación. Entonces, una misma conversación podrá tener una interpretación distinta si se varía su contexto de análisis.

Interpretación = *Representación abstracta* + *Contexto*

4.1.1 Universo de temas

El contexto de conocimiento determina cuales son todos los temas que se puedan llegar a tener en cuenta, tanto en el momento de realizarse la conversación, como cuando ésta es analizada.
Además, el desarrollo de una conversación puede asociarse a un conocimiento implícito que se tiene en el momento que se realiza. El mismo variará dependiendo de este contexto de conocimiento en el que la misma se desarrolle o sea analizada. Para clarificar la idea anterior, veamos los siguientes ejemplos.
El entorno de un niño de dos años (Ignacio) tiene un conjunto de temáticas al cual vincula todas sus expresiones y las cosas que escucha. Estas temáticas pueden decirse que son, a grandes rasgos: la familia, la comida, los juegos y su cuerpo. Este es el entorno del niño. Este entorno tiene asociado conocimiento implícito, y conocimiento explícito, que determinará en gran parte su modo de expresarse y las palabras que use. Cada palabra o frase que escuche la asociará solamente a alguno de estos grupos descriptos, porque no

conoce de la existencia de otras cosas. Y es poco probable que haga referencia a cosas o situaciones en forma implícita. En cambio sí habrá una tendencia a repetir conceptos, es decir, usar redundancias al expresarse, en la forma en que un adulto no lo haría.
Por otra parte, en una comisión de médicos hablando de temas de su profesión, se emplearán palabras específicas que no serán expresadas, porque se tiene un conocimiento implícito de las mismas y se presupone que los demás interlocutores también lo tienen.
A su vez, estos médicos, para referirse a un tema en particular, necesitarán expresar una cantidad considerable de palabras para que los demás interlocutores lo relacionen a ese tema en particular por sobre los demás temas referidos a la medicina. Y mayor será el esfuerzo realizado para ser comprendido (mayor la cantidad de palabras y frases específicas del tema) mientras más específico sea el tema en cuestión.
Resumiendo, las palabras que se encuentren en una conversación, y por ende las palabras necesarias para la identificación del tema, dependerán del conocimiento implícito que manejen sus interlocutores, y del conocimiento explícito necesario que utilicen para la intercomunicación, dentro del contexto de conocimiento en el cual la conversación se desarrolla. De aquí se deduce que la cantidad y calidad de los temas considerados como el total de temas posibles determinará características en la conversación. Y estas características deberán ser tenidas en cuenta para deducir el tema.
Entonces, para el análisis de una conversación, será necesario en primer lugar, decidir cual es el universo de temas en el que trabajará la herramienta. Este universo determinará muchos factores a considerar, como por ejemplo qué palabras serán destacadas en una conversación, cuáles no aportarán información, qué tan importantes son, qué cantidad de palabras será necesaria para poder reconocer un tema, y qué cantidad de palabras será necesaria para diferenciar un tema de otro.

Definición 4.1: [Aporte Semántico de las palabras en la conversación]
Aporte Semántico (**AS**) de una palabra es una medida de la cantidad de información que brinda esa palabra y que sirve para orientar en la búsqueda del tema de una conversación.
Para la valuación de cada palabra, se definen las siguientes condiciones:

1. Un **AS** es mayor cuanto menor es la probabilidad de ocurrencia de esa palabra en una conversación obtenida al azar, y menor en caso contrario.

También debe considerarse que,

2. Un **AS** es mayor cuanto mayor es la probabilidad de ocurrencia de una palabra en un tema específico o dentro de una familia de temas, y menor en caso contrario.

Con la condición (2) se estará asignando un bajo valor a palabras que, teniendo una baja probabilidad de aparición en una conversación tomada al azar, aporten muy poca información, como ser palabras caídas en desuso, tal como sucede en Argentina con el verbo "yacer".

A partir de las condiciones (1) y (2) se puede construir una función que asigne un valor a cada palabra dentro del idioma (en este caso castellano), y ordenar las palabras de mayor a menor según esta valuación. Con este ordenamiento, las palabras destacadas de la

conversación serían las primeras *n* palabras para algún *n* dado en forma arbitraria (más adelante formalizaremos estas definiciones).
El inconveniente que presenta esta forma de valuación es que por cada palabra, es necesario calcular el aporte semántico, y esto implica analizar cuál es la probabilidad de aparición de una palabra en una conversación referida a un tema, teniendo en cuenta el universo de todos los temas existentes.
Aunque en teoría esto sería ideal, en la práctica es algo difícil de realizar.

Ejemplo 4.1:
La palabra "estar" (verbo) tiene un AS muy bajo, debido a que la probabilidad de que se mencione (el verbo en infinitivo o algún derivado del mismo) en una conversación tomada al azar es muy alta. Entonces esta palabra no aporta información a partir de la cual se pueda deducir el tema de conversación.
La palabra "Maradona", en cambio, tiene un AS muy alto, ya que es muy poco probable que aparezca en una conversación tomada al azar. Pero en cambio es ciertamente probable que sí aparezca en conversaciones relacionadas a "Fútbol", "Farándula" o "Adicciones".

4.1.1.1 Discretización del AS

Una variante a la estrategia anterior es discretizar los valores asignados a cada palabra, asumiendo que si una palabra no aporta información para ninguna conversación del universo de temas en cuestión (palabra o frase descartada) entonces vale 0; si tiene un valor de distinción en un tema relacionado entonces vale 1; o que tiene un valor especial para un tema, denominándola en este caso clave, y asignándole un valor mucho mayor a 1. Gracias a esta discretización se obtiene un esquema de valuación más simplificado. La determinación de las palabras descartadas, palabras destacadas, y claves dependerá del universo de temas definido. También se incluirán a las frases en esta valuación y el ordenamiento.
La opción que adoptamos para determinar el aporte semántico de las palabras o frases es la *discretización.*

4.1.2 Patrones de Comportamiento en la conversación

Patrón de Comportamiento de una palabra o frase es una característica que la misma tiene, depende de todas las veces que se la mencione en la conversación y brinda información necesaria para la deducción del tema.

Definición 4.2: [Ocurrencia de la palabra o frase]
Ocurrencia de una palabra o frase es cada mención en la conversación, de esa palabra o frase.

A partir de la definición 4.2. se asume que toda palabra en una conversación tiene asociado un conjunto de ocurrencias. A su vez cada ocurrencia tiene características relativas al lugar en la conversación donde se mencionó, como ser el usuario, el párrafo y la posición dentro del párrafo. Se necesita una métrica que exprese cuando una palabra o frase es relevante dentro de la conversación. Para esto definimos el *peso de una palabra o frase.*

Definición 4.3: [Peso de una palabra o frase]
Peso de una palabra o frase es el valor dado por una función dependiente de los valores vinculados a cada uno de los patrones de comportamiento asociados a estas palabras. La idea es pesar las palabras de tal forma que un mayor peso represente una mayor importancia de esa palabra, y así poder seleccionar un subconjunto relevante del total de palabras existentes en la conversación formado por las palabras de mayor peso. *Se calcula el peso como el producto de cada uno de los factores asociados a los patrones de comportamiento.*

Observación 4.1:
Para que cada factor del producto definido en 4.3 influya de una manera controlada sobre el resultado calculado, es necesario normalizar cada uno de estos valores entre una cota mínima y una cota máxima preestablecidas.

Existen distintos factores asociados a los patrones de comportamiento, que son los siguientes:

1. Distancia Promedio.
2. Frecuencia.
3. Concentración.
4. Experiencia del Usuario.
5. Condición de Palabra Clave.
6. Necesidad de una palabra.

4.1.2.1 Distancia promedio

Para tener una medida de cómo está distribuida una palabra en el contexto de la conversación. Para esto se define *distancia entre párrafos* como la cantidad de párrafos existentes entre un par elegido en la conversación. Esta definición se extiende a las palabras que están en dos párrafos elegidos de la conversación. Si dos palabras están en el mismo párrafo, su distancia es 0.
La *distancia promedio* de una palabra (o frase) es la suma de las distancias de cada palabra con todas las demás palabras o frases, dividido la cantidad total de palabras y frases. Cuanto mayor sea la distancia promedio, menos debería influir en el peso de la palabra. Entonces el valor que influirá en el peso es 1/*distancia promedio*.

4.1.2.2 Frecuencia

Una palabra o frase tendrá tantas ocurrencias como veces se la mencione en la conversación.
La *frecuencia* es la cantidad de ocurrencias de la palabra, dividido la cantidad total de ocurrencias de la conversación.

4.1.2.3 Concentración

La *concentración* es una medida que expresa la continuidad de una misma palabra a través de los distintos párrafos de una conversación. Lo que se pretende lograr es una medida que exprese cuándo una palabra sirve de estímulo en la conversación. De tal forma, se considerará que una palabra que se repite de manera muy seguida es un estímulo importante para el contexto en el cual se repite. Pero si una palabra que quizás tenga una frecuencia considerable, está localizada en una sección acotada de la conversación (una alta concentración), no deberá ser muy importante, ya que la palabra se supone que fue usada en ese contexto específico de la conversación, y no en la conversación en general.
Entonces, para una misma cantidad de ocurrencias, *una baja concentración da la idea de que la palabra fue usada en toda la conversación, independientemente de cualquier contexto específico.*
Para calcular la concentración, se debe analizar cada cuanto se repite una palabra o frase a partir del momento en el que se la menciona por primera vez. A partir de la primera aparición de una ocurrencia de palabra en un párrafo, se quiere saber cuál es la siguiente aparición de una ocurrencia de la misma palabra en un párrafo distinto. Hasta aquí se considera la distancia entre una ocurrencia y la siguiente. La idea es observar esta distancia a través de las distintas ocurrencias (es decir la distancia de la última ocurrencia tenida en cuenta con la aparición de la siguiente a ésta), y sumar cada una de estas distancias obtenidas. Si se divide esta suma por la cantidad de ocurrencias, estamos obteniendo una medida del promedio de espera de la siguiente aparición. Entonces, si tenemos por ejemplo dos apariciones, en el párrafo 1 y 2, o tres apariciones, en los párrafos 1, 2 y 3, la medida obtenida sería en ambos casos 1. Pero en realidad es más importante la segunda opción, donde se ve que una palabra aparece con una mayor frecuencia e igualmente seguido, con lo cual debemos agregar un factor adicional a este cálculo que nos exprese esta preferencia. Esto se consigue dividiendo el factor obtenido por la cantidad de apariciones de la palabra. Entonces para el primer caso daría 1/2, y para el segundo 1/3, donde se expresan claramente cada una de las distintas situaciones. Entonces el valor que influirá en el peso es 1 / *concentración*.

4.1.2.4 Experiencia del usuario

Se quiere dar una relevancia especial a palabras dichas por un usuario, siempre y cuando estas palabras estén vinculadas al tema en el cual el usuario es experto. Es decir, se quiere inferir que hay un considerable aumento de la probabilidad de que un usuario experto en un tema esté hablando del mismo, si se está ante la presencia de una conversación en la cual dicho usuario nombra palabras relacionadas al tema.
Para aprovechar esta información en la etapa de la interpretación, hay que tener en cuenta que estas palabras especiales pueden ser dichas por distintos usuarios, sean expertos o no en uno o varios temas a los cuales la palabra esté ligada.
Entonces, la *experiencia del usuario* en la etapa de la interpretación es el promedio del grado de experiencia de los usuarios que mencionan cada palabra en la conversación, palabras vinculadas a los temas que incluyan esa palabra.
Para todas las palabras y frases, al leer cada ocurrencia se le asocia un peso que es representativo de la experiencia del usuario, siempre que esa palabra o frase sea una palabra distinguida en algún tema en el cual el usuario se considere experimentado. Luego el peso de usuario relacionado a dicha palabra será el promedio de todos estos pesos.

Aparente desventaja: puede ser que esta palabra también forme parte de otro tema que no tiene nada que ver con la experiencia del usuario, ni la conversación analizada.
Ventaja: es de esperarse que el resto de las palabras que identifiquen a este otro tema no vinculado a la experiencia del usuario sean muy distintas.
Entonces se dice que la probabilidad de que el mismo subconjunto de palabras con un peso especial de usuario esté en otro tema distinto será muy baja, lo cual asegura estar favoreciendo solamente al acercamiento del tema en el cual el usuario es experto.

Observación 4.2: Si un usuario es "experto", con un cierto grado (peso) en un tema; entonces, en la base de conocimiento se almacena el usuario, el tema en el cual es experto, *y un valor entre 1 y 4 fijado arbitrariamente, que representa el grado de experiencia que tiene el usuario en este tema*. De igual forma se procede con cada tema en el que el usuario es experto.

4.1.2.5 Condición de Clave

De la conversación se extraen las palabras y frases de mayor relevancia sobre la base del cálculo de los patrones, y las palabras o frases que figuren como claves en la base de datos (Sección 3.1). *El valor asociado a una clave es un valor entre 1 y 4 que se establece en forma arbitraria para cada clave, de modo tal que represente la incidencia que tenga la misma en la determinación del peso de la palabra o frase.*

4.1.2.6 Necesidad de una Palabra

Hay palabras en una conversación que se distinguen por ser mencionadas frecuentemente. Tal vez hasta el punto de caer en la redundancia. Pero es notable que para expresar las ideas es necesario nombrarlas de manera casi constante. Este tipo de palabras estará caracterizado por tener una alta frecuencia, y además estar distribuido de manera uniforme a lo largo de la conversación. Es decir, son palabras que tienen una alta frecuencia y una baja concentración, relativa a esta frecuencia.
Por otra parte, una palabra que se distinga por una alta frecuencia, pero que también tenga una alta concentración, se dirá que es *necesaria* en un contexto de la conversación.
Como esta medida depende directamente de dos valores que ya son calculados y analizados en la conversación, no se analizará en forma directa, sin que esto signifique despreciar la importancia que la misma tiene como patrón determinante en la conversación.

4.1.3 Grafo de la conversación

Sea $G_C(V, X)$ un grafo en el cual V es el conjunto de palabras de una conversación y X es un conjunto de pares no ordenados formados por 2 palabras distintas de la conversación. Luego G_C se denomina *grafo de la conversación*.

Definición 4.4: [Distancia promedio entre un par de palabras o frases]
La *distancia promedio entre un par de palabras o frases* es el promedio de las distancias entre cada ocurrencia de una de ellas y cada ocurrencia de la otra.

Con la información obtenida a partir del cálculo de cada uno de los patrones de las palabras y frases leídas, se construye un grafo que contiene la representación abstracta de la

conversación original. Cada nodo representará a cada una de las palabras presentes en la conversación, y tendrá asociado el peso calculado para esa palabra o frase. Cada arco tendrá una medida que será la *distancia promedio* entre ese par de palabras o frase que vincula.

4.1.4 Resultado de la Interpretación

El objetivo de la interpretación es distinguir la información más relevante de la conversación, y descartar las restantes. Para ello, debe:
a) Podar el grafo que representa la conversación.
b) Seleccionar los nodos con mayor peso, del grafo podado.
El proceso de poda consiste en descartar las palabras o frases que tienen alta distancia promedio, baja frecuencia, y alta concentración, de acuerdo a cotas preestablecidas. Se interpreta que una palabra o frase que tiene alta distancia promedio está muy alejada del resto de la conversación y por eso no será tenida en cuenta. Si tiene baja frecuencia, no es muy importante y aporta poca información, y si tiene alta concentración se interpreta que la palabra o frase sólo es necesaria en un contexto específico de la conversación

Las palabras claves no se descartan.

Definición 4.5: [Palabras destacadas de la conversación]
Las *palabras destacadas de la conversación* son el conjunto de palabras elegidas como resultado de la interpretación luego de podar el grafo y seleccionar las palabras de mayor peso.

Definición 4.6: [Máxima cantidad de palabras destacadas de la conversación]
La *máxima cantidad de palabras destacadas de la conversación* es una cota que determina el mayor número de palabras elegido del grafo de la conversación.

El proceso de selección de las palabras destacadas consiste en ordenar, según sus pesos, las palabras que han quedado representadas en el grafo podado, y luego elegir las *n* palabras de mayor peso según el máximo establecido (def. 4.6). A partir de este conjunto de palabras seleccionadas, y de las palabras claves, se buscará el tema asociado a la conversación.

4.2 Descubrimiento de Temas

Con la información contenida en el grafo podado, producto de la interpretación, se buscarán los temas que tengan coincidencias con el resultado de esta representación abstracta obtenida. En teoría, para la búsqueda de un tema, sería suficiente con recorrer la base de temas, eligiendo el tema que tenga más palabras asociadas a la conversación. De esta manera se estaría eligiendo el tema más "parecido" a la conversación original.
Esta estrategia presenta dos inconvenientes. El primero y principal, la búsqueda sería costosa, en la medida que la base contenga una considerable cantidad de temas almacenados, y se tenga una gran cantidad de palabras de la conversación. La segunda es que, al carecer de un análisis de la conversación, no se tiene una forma de poder inferir nuevos temas, asociados o no a los ya existentes en la base de conocimiento, con lo cual

no se considera la posibilidad de relacionar distintos temas entre sí, ni tampoco la capacidad de aprendizaje a partir de información ya obtenida.
En cambio, la estrategia a seguir se basará en el reconocimiento de un subconjunto de palabras y frases representativo de la conversación, y encontrar el tema que mejor se aproxime a este conjunto de palabras, con la posibilidad de comparar contra conjuntos similares (de palabras identificadoras de temas) ya existentes.

4.2.1 Tema candidato

Sobre el conocimiento que se tiene de los temas almacenados y la información de la conversación analizada, se realizará una estrategia de selección. Obviamente las estrategias que pueden pensarse son muy variadas. Para el presente trabajo se desarrolló la estrategia que se explica a continuación, porque es la que maneja toda la información obtenida de la conversación, y la existente en la base de conocimiento; cada puntuación obtenida es una medida que refleja cada una de las características observadas en la conversación, y la relación de la misma con el tema a comparar.
Se calculará un puntaje para cada tema. Este puntaje es interpretado como una medida de proximidad (o similitud) entre cada tema y el conjunto de palabras y frases representativas de la conversación.
El conjunto de palabras que se usará para la búsqueda del tema, es el resultante del proceso de interpretación de la herramienta (ver Sección 4.1.4). El tema que tenga el puntaje más alto será el *tema candidato*.

Para el cálculo del puntaje se tendrá en cuenta:

- *Peso de la palabra en el tema*: corresponde 1 si la palabra no es clave, y el peso de la clave, si lo es. En la base de conocimiento se almacena cada palabra clave, el tema del cual es clave y el peso que tiene esa clave en el tema.
- *Experiencia del usuario* que expresa la palabra en la conversación. Si un usuario menciona una palabra relacionada a un tema en el cual es experto, es de esperar que se le dé un puntaje especial al momento de calcular el puntaje de dicho tema. En la base de conocimiento se almacena cada usuario experto, el tema del cual es experto y el peso que tiene ese usuario experto en el tema.
- *Palabras Coincidentes*: palabras que están en el tema y en la conversación.
- *Palabras Faltantes*: palabras que están en el tema pero no en la conversación.
- *Palabras Sobrantes*: palabras que están en la conversación y no en el tema.
- *Coincidentes*: es la suma de los pesos de las palabras coincidentes
- *Faltantes*: es la suma de los pesos de las palabras faltantes
- *Sobrantes*: es la cantidad de palabras sobrantes

A partir de estas definiciones, el puntaje de un tema será

- puntajetema = coincidentes / (coincidentes + faltantes)

El objetivo de la búsqueda es encontrar en la base de conocimiento, el tema candidato de la conversación. Para ello, selecciona el tema con mayor puntaje, obtenido de acuerdo a los cálculos explicados en los párrafos anteriores.

4.3 Resumen

En el presente capítulo se presentó la base teórica y definiciones necesarias para distinguir la información relevante de una conversación, y que la misma sea suficiente para construir una representación abstracta de ésta. Entonces se definió el aporte semántico de los distintos elementos de la conversación y los patrones de comportamiento de las palabras y frases presentes. Se describió el diseño de una estrategia de búsqueda y comparación entre la representación abstracta de la conversación y los temas almacenados en la base de conocimiento.

La definición de una interpretación adecuada es el punto crítico de la aplicación. De la información tomada para crear la representación abstracta del conocimiento contenido en una conversación dependerá el éxito que tenga la aplicación para deducir correctamente el tema o la relación con un tema. Es decir, de esto dependerá que la aplicación cumpla con el fin para el cual fue creada.

Existe una relación directa entre la forma en que se representa abstractamente una conversación y la información que es almacenada para cada tema en la base de conocimiento. Ambas representaciones (la de la conversación y la de los temas) deben permitir que sean comparables entre sí y que como resultado de dicha comparación se deduzca la relación semántica que exista entre ambas.

En el siguiente capítulo se verá para el tema encontrado, que cumple con la propiedad de ser el más "parecido" a la conversación, qué tipo de relación se establece entre la conversación y este tema, dependiendo de que tan parecido sea, y cual es la decisión adecuada que deba tomar la herramienta frente a la situación existente.

Capítulo 5

Decisión de Resultados

A partir del tema elegido en la búsqueda de temas relacionados a la conversación es necesario decidir cuál es la relación que guarda con la conversación analizada.

5.1 Selección del tema de la conversación

Para la elección de un tema se analizan los siguientes factores:
1) Cantidad mínima de palabras asociadas al tema
2) Coincidentes, faltantes y sobrantes

5.1.1 Cantidad mínima de palabras asociadas al tema

Cuando no se encuentra un tema (cantidad de coincidentes = 0), tenemos que determinar qué hacer con las *palabras destacadas de la conversación* (def. 4.5). Si esas palabras son muy pocas, se asumirá que las mismas no aportan la información mínima necesaria para deducir a partir de éstas un *nuevo* tema. Entonces necesitamos un parámetro de comparación para determinar en forma automática este mínimo de palabras necesarias para deducir un tema.

Definición 5.1: [Mínimo de Palabras en un tema]
Mínimo de Palabras en un tema (MPT) es la menor cantidad de palabras necesarias para identificar un tema.

Definición 5.2: [Distancia entre temas]
Distancia entre dos temas distintos es la mayor cantidad de palabras identificadoras de uno de los temas y que no son identificadoras del otro.

Ejemplo 5.1:
Tema A "Casamiento".
Palabras asociadas: "Torta", "Anillo", "Novio", "Baile"
Tema B "Cocina".
Palabras asociadas: "Torta", "Cocinar", "Amasar", "Harina", "Sal", "Aceite", "Condimento"

Como A tiene 3 palabras que B no tiene, y B tiene 6 palabras que A no tiene, la Distancia entre el tema A y B es 6.

Definición 5.3: [Distancia Mínima entre temas]
Distancia Mínima entre temas (DMT) es la menor distancia que existe entre dos temas cualquiera de la base de conocimiento.

Definición 5.4: [tema nulo]
Tema Nulo es el tema sin palabras identificadoras

5.1.1.1 Relación entre *MPT* y *DMT*

La mínima distancia entre cualquier tema y el tema nulo (*DMT)* es el *MPT*, ya que si hubiera un tema con una distancia menor al tema nulo, no cumpliría la condición del *MPT*. Por lo tanto *MPT* y *DMT* se refieren a un mismo valor.
Estas nuevas definiciones (*MPT* y *DMT*) permiten determinar si un conjunto de palabras seleccionadas por su importancia, en la conversación puede estar asociadas a un tema (si el tamaño del conjunto es mayor o igual a *MPT*), o deducir que hay poca información para poder asociar las palabras que se tienen con algún tema (si el tamaño del conjunto es menor a *MPT*).

5.1.2 Coincidentes, Faltantes y Sobrantes

Además de la cantidad mínima de palabras asociadas a un tema (Sección 5.1.1), se analizan los valores de *coincidentes, faltantes* y *sobrantes*, respecto a umbrales o *cotas de rechazo, incertidumbre y aceptación* (la definición de coincidentes, faltantes y sobrantes se explica en la Sección 4.2.1, y la explicación de umbrales o cotas se da en la Sección 5.2.2).

Una aproximación ideal a un tema tiene asociado un alto valor de coincidentes y bajos valores de faltantes (alto puntaje del tema).
Además se desea un bajo valor de sobrantes, que determinará que sólo una pequeña parte de la conversación no está fuertemente ligada al tema elegido.
Por otra parte, un alto valor de sobrantes indicará que una gran parte de la conversación no está siendo contemplada por el tema elegido. Esto combinado con el hecho de que el tema elegido tiene un alto grado de coincidentes, significará que la conversación se refiere a un subtema del tema elegido, y que será determinado a partir de las palabras sobrantes.
El puntaje para decidir la creación de un subtema estará asociado a un alto valor de sobrantes y de coincidentes con el tema padre. Si el subtema ya existe entonces tendrá un alto valor de coincidentes y un bajo valor sobrantes.

Observación 5.1: Los valores de coincidentes y faltantes se resumen en un único parámetro que es el *puntaje del tema* (ver Sección 4.2.1).
Se observa que un alto puntaje del tema significa un alto valor de coincidentes y un bajo valor de faltantes; en cambio un bajo puntaje del tema significa un bajo valor de coincidentes y un alto valor de faltantes. Los valores de coincidentes y faltantes están inversamente relacionados.

5.2 Casos posibles para tomar una Decisión

El objetivo de la búsqueda es encontrar en la base de conocimiento, el tema candidato de la conversación (Sección 4.2.1). Los resultados de esta búsqueda pueden ser:

1) No se encontró un *tema candidato*.
2) Se encontró un *tema candidato* de la conversación.

5.2.1 No se encontró un tema candidato

Si no se ha encontrado un tema candidato, se analizan las palabras destacadas (es decir, con mayor peso) en la conversación.

- Si la cantidad de palabras destacadas de la conversación es muy baja (menor a *MPT*), se le informa al usuario que no se ha encontrado ningún tema.
- En cambio, si ésta cantidad de palabras es alto (mayor a *MPT*), las mismas son mostradas al usuario, y se le pide que informe el nombre del tema de la conversación.

5.2.2 Se encontró un tema candidato

Si se encontró un tema candidato, se analizan los *sobrantes y puntaje* del mismo (Sección 4.2.1) de acuerdo al lugar que ocupan en los intervalos de estudio de la Figura 5.1.

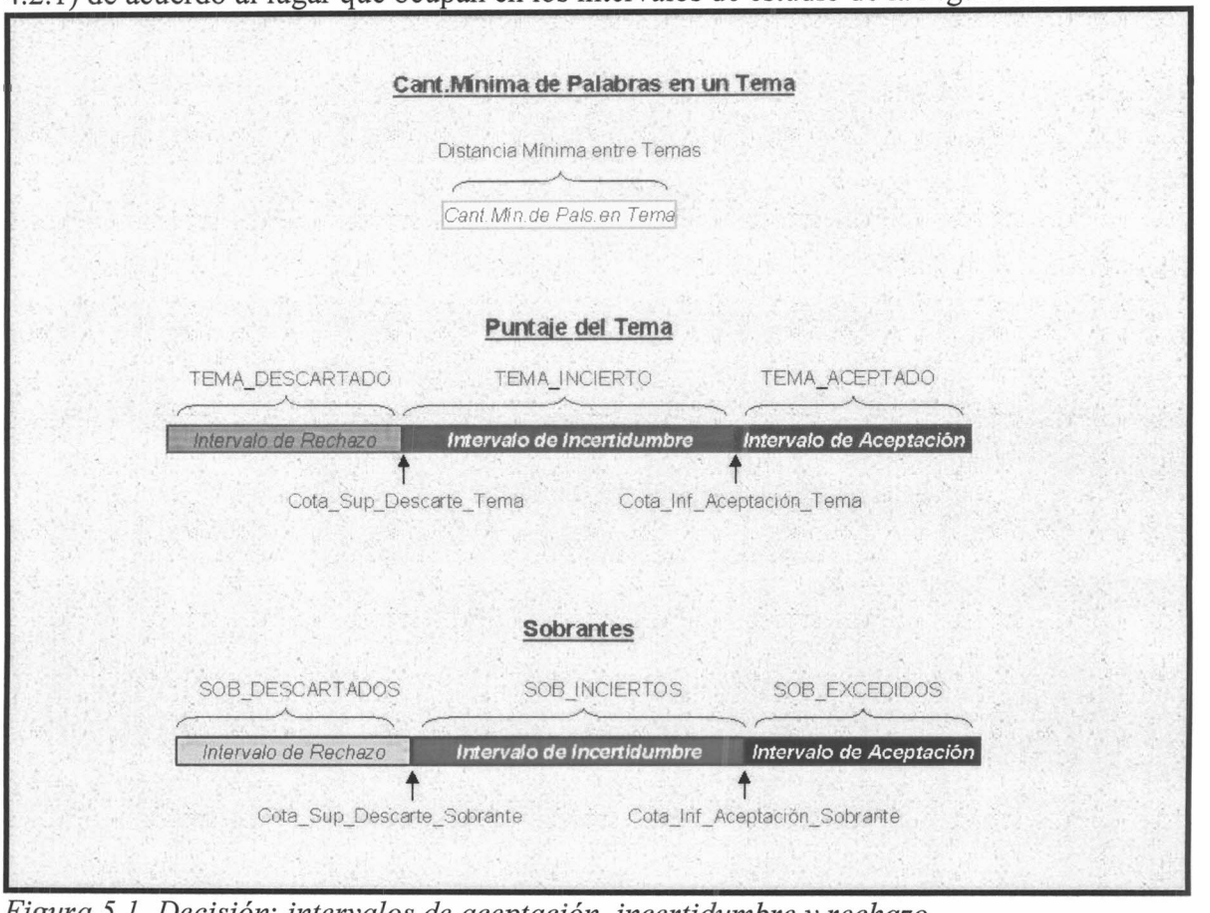


Figura 5.1. Decisión: intervalos de aceptación, incertidumbre y rechazo.

5.2.2.1 Intervalos de estudio

Definición 5.5: [Intervalo de estudio]
Intervalo de estudio es un rango de valores determinados por un límite inferior y un límite superior. Estos límites se llaman *cotas* o *umbrales*.

Los tipos de intervalos de estudio son:
- Intervalo de rechazo
- Intervalo de incertidumbre
- Intervalo de aceptación

Los tipos de cotas son:
- Cota superior de descarte.
- Cota inferior de aceptación.

Los valores del intervalo de rechazo son menores a la cota superior de descarte.
Los valores del intervalo de incertidumbre son mayores a la cota superior de descarte y menores a la cota inferior de aceptación.
Los valores del intervalo de aceptación son mayores a la cota inferior de aceptación.

5.2.2.2 Análisis del puntaje del tema y los sobrantes

Los intervalos de estudio se usan para analizar:
1) El puntaje del tema candidato (*pt*).
2) Los sobrantes (*sob*).

- Tema descartado
  Si *pt* se ubica en el intervalo de descarte entonces el tema candidato se descarta.
- Tema incierto
  Si *pt* se ubica en el intervalo de incertidumbre entonces la elección del tema candidato es incierta o dudosa.
- Tema aceptado
  Si *pt* se ubica en el intervalo de aceptación entonces el tema candidato o un subtema es aceptado como *tema de la conversación*.

- Sobrantes descartados
  Si *sob* se ubica en el intervalo de descarte entonces las palabras sobrantes se descartan para el análisis de un segundo tema de la conversación.
- Sobrantes inciertos
  Si *sob* se ubica en el intervalo de incertidumbre entonces la influencia de las palabras sobrantes en algún tema es incierta.
- Sobrantes excedidos
  Si *sob* se ubica en el intervalo de aceptación entonces la cantidad de palabras sobrantes es relevante, y se tienen en cuenta para la identificación de un segundo tema de la conversación.

5.2.2.3 Casos de Decisión

Al analizar los valores del puntaje del tema candidato y los sobrantes en forma conjunta, surgen 9 casos posibles para luego tomar una decisión:

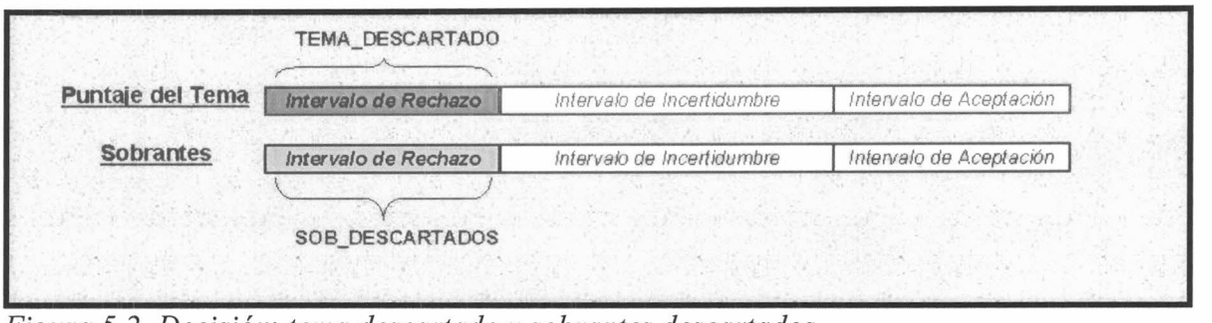


Figura 5.2. Decisión: tema descartado y sobrantes descartados.

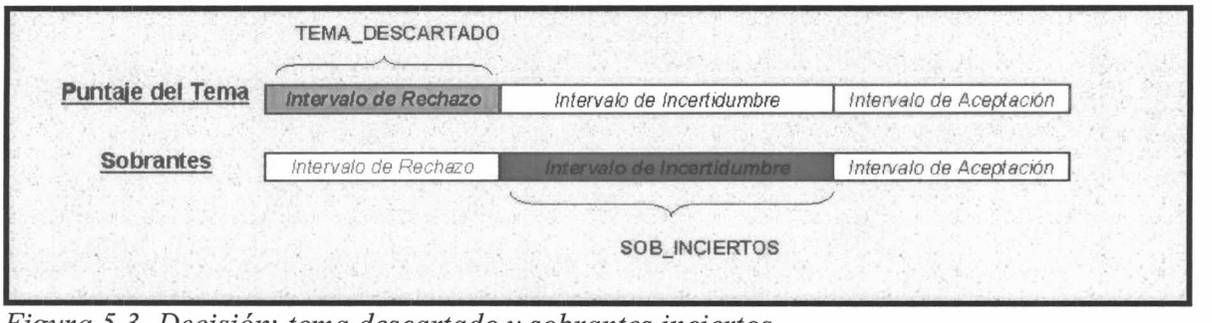


Figura 5.3. Decisión: tema descartado y sobrantes inciertos.

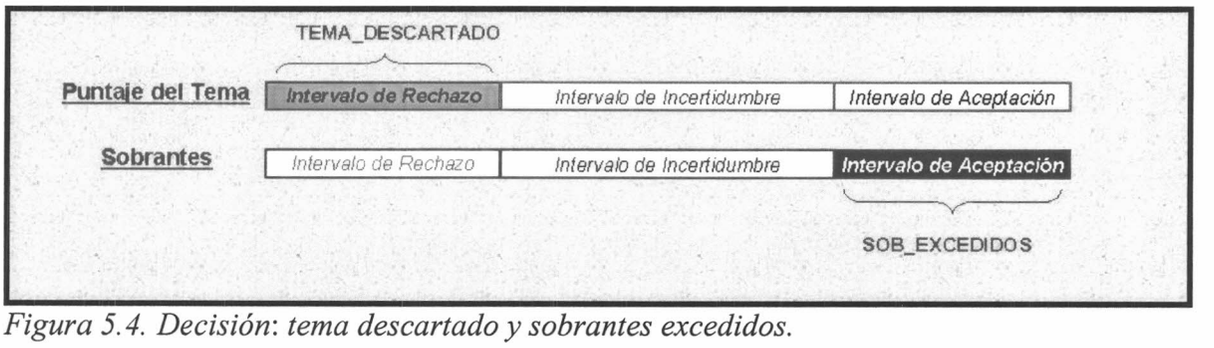


Figura 5.4. Decisión: tema descartado y sobrantes excedidos.

En los casos en que el tema se considera descartado (Figuras 5.2, 5.3 y 5.4) se tiene poca información asociada al tema encontrado (tema candidato). Si la información de la conversación (cantidad de palabras coincidentes o sobrantes) es relevante (mayor a *MPT*), se utiliza para que el usuario determine el tema. En caso contrario, sólo se informa al usuario que hay poca información para relacionarla a un tema.

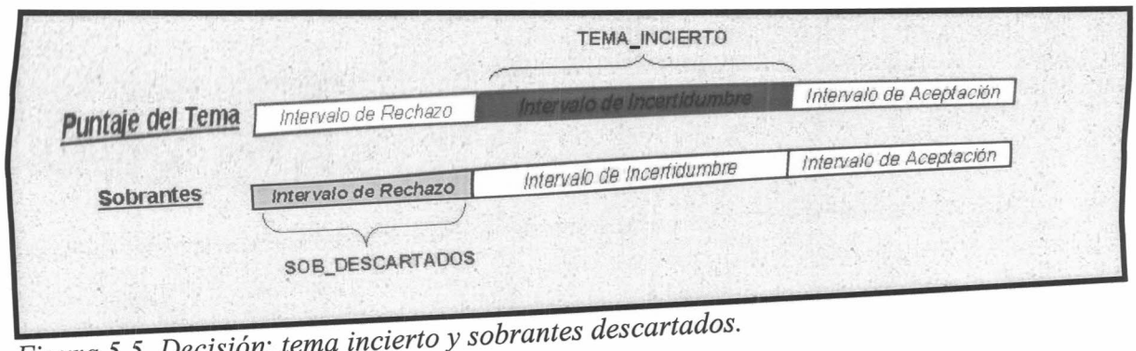


Figura 5.5. Decisión: tema incierto y sobrantes descartados.

Tema incierto y sobrantes descartados. El puntaje del tema candidato se encuentra en el intervalo de incertidumbre, entonces las palabras faltantes pueden diferenciar al tema padre del tema hijo. El tema padre es el tema de la conversación, y el *tema hijo* (subtema) es el tema encontrado en la base de conocimiento (tema candidato). Si la cantidad de faltantes es relevante (mayor a *MPT*), se utiliza para que el usuario determine el tema de la conversación, el cual será el tema padre del tema candidato. En caso contrario, si la cantidad de coincidentes es relevante (mayor a *MPT*), la herramienta afirma que se está hablando del tema candidato, sino sólo se informa al usuario que hay poca información para relacionarla a un tema. La cantidad de sobrantes está en el intervalo de rechazo, entonces no son suficientes para el análisis de un segundo tema en la conversación.

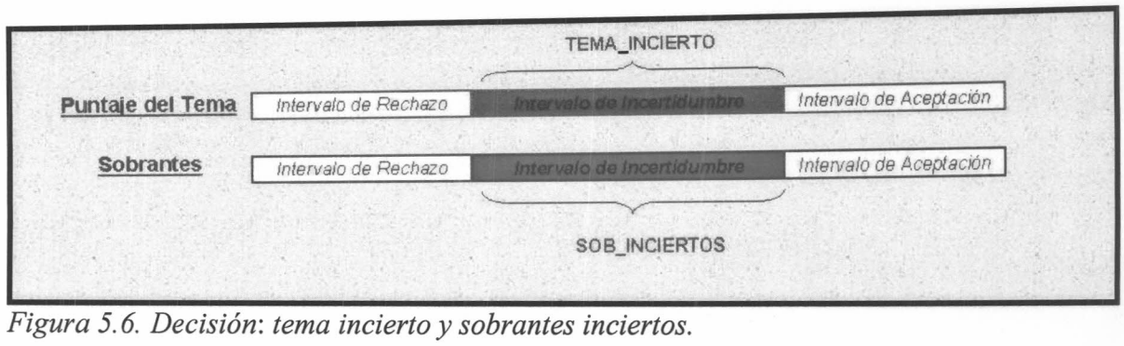


Figura 5.6. Decisión: tema incierto y sobrantes inciertos.

Tema incierto y sobrantes inciertos. El puntaje del tema candidato se encuentra en el intervalo de incertidumbre y la cantidad de sobrantes también se encuentra en el intervalo de incertidumbre. Es probable que se esté hablando del tema encontrado en la base de conocimiento (si la cantidad de faltantes no es relevante) o de un tema padre del tema encontrado (si la cantidad de faltantes es relevante). También es posible que se esté hablando de un segundo tema, determinado por una cantidad considerable de palabras sobrantes.

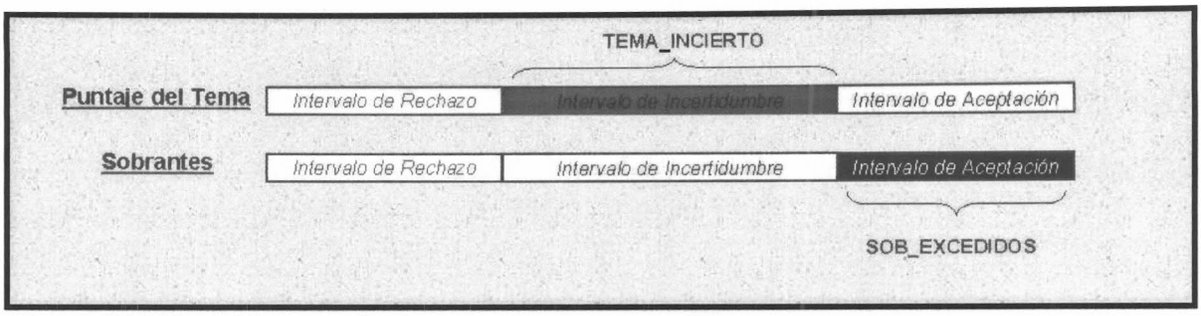


Figura 5.7. Decisión: tema incierto y sobrantes excedidos

Tema incierto y sobrantes excedidos. La cantidad de sobrantes es importante, porque se encuentra en el intervalo de aceptación. Esto permite deducir que las palabras sobrantes están identificando un tema de la conversación.
Es probable que se esté hablando del tema candidato (si la cantidad de faltantes no es relevante) o de un tema padre del tema candidato (si la cantidad de faltantes es relevante). También se está hablando de un segundo tema, determinado por una cantidad importante de palabras sobrantes.

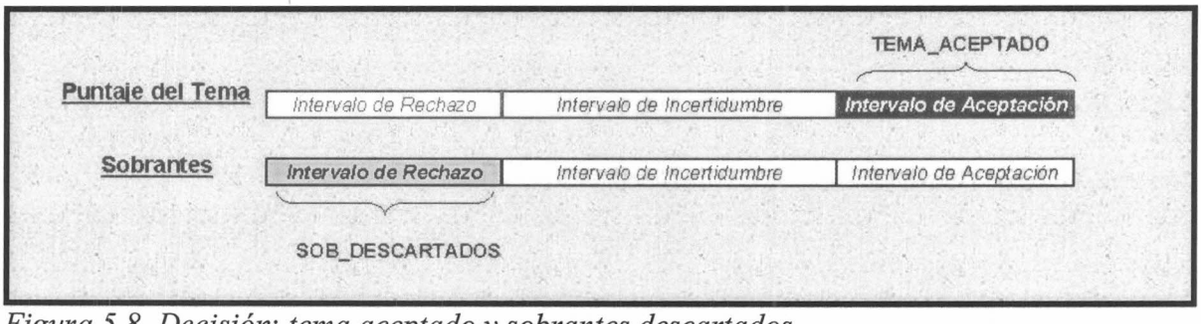


Figura 5.8. Decisión: tema aceptado y sobrantes descartados

Tema aceptado y sobrantes descartados. Este es el caso óptimo. Se está hablando de un tema de conversación en forma exclusiva (el tema candidato). Se le informa al usuario el nombre del mismo, y las palabras que lo identifican.

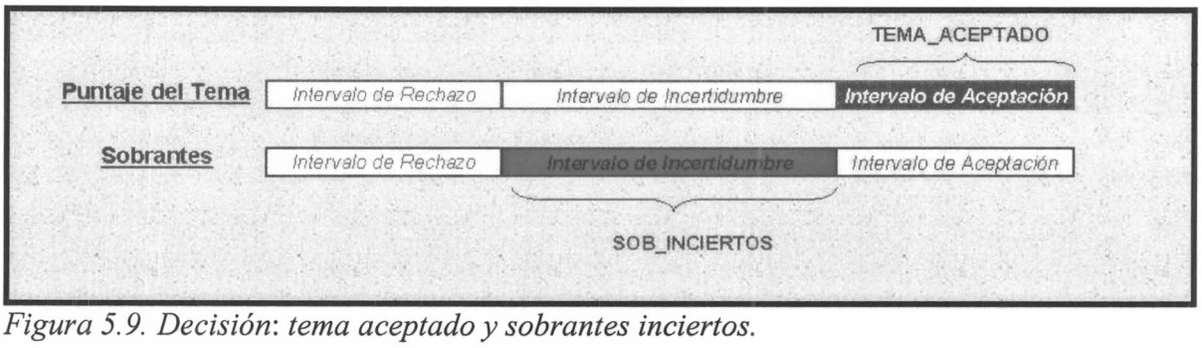


Figura 5.9. Decisión: tema aceptado y sobrantes inciertos.

Tema aceptado y sobrantes inciertos. La cantidad de sobrantes se encuentra en el intervalo de incertidumbre, entonces las palabras sobrantes pueden diferenciar al tema hijo (subtema)

del tema padre. El tema hijo es el tema de la conversación, y el tema padre es el tema encontrado en la base de conocimiento (tema candidato). Si la cantidad de sobrantes es relevante (mayor a *MPT*), se utiliza para que el usuario determine el tema de la conversación, el cual será un subtema (o tema hijo) del tema candidato. En caso contrario, si la cantidad de coincidentes es relevante (mayor a *MPT*), la herramienta afirma que se está hablando del tema candidato, sino sólo se informa al usuario que hay poca información para relacionarla a un tema.

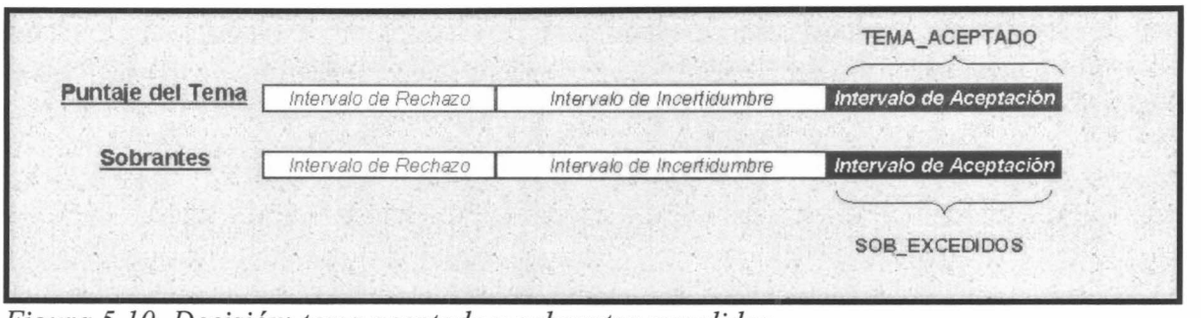


Figura 5.10. Decisión: tema aceptado y sobrantes excedidos.

Tema aceptado y sobrantes excedidos. El puntaje del tema se encuentra en el intervalo de aceptación y la cantidad de sobrantes también se encuentra en el intervalo de aceptación.
El alto puntaje del tema determina que el tema de la conversación es el tema candidato. Pero la importante cantidad de sobrantes puede determinar un fraude en la conversación.
La decisión que toma la herramienta es informar al usuario el nombre del tema que se encontró, y la posible existencia de un fraude.

5.3 Resumen

En el presente capítulo se presentó un estudio de todas las situaciones posibles con las que se puede encontrar la aplicación en el momento de decidir qué hacer si se encontró un tema, si no se encontró, si es necesaria el aporte del usuario o si sólo se le informarán los resultados obtenidos.
La definición de los intervalos de incertidumbre es determinante no sólo para el resultado inmediato ofrecido al usuario, sino que, en consecuencia, también tiene aparejado la incorporación de nuevo conocimiento, y del tipo de conocimiento que es inferido.
A partir de la cantidad mínima de palabras en tema se deduce cuantas palabras son necesarias para diferenciar un tema padre y un hijo, dos temas distintos, y la existencia de un mínimo de información necesaria para deducir un tema.
En el siguiente capítulo se verá la funcionalidad de cada una de las componentes de la aplicación. También se explicarán las distintas situaciones que permiten que la aplicación desarrolle un autoaprendizaje.

Capítulo 6

Arquitectura

Una vez presentado el marco teórico sobre el que está basado el diseño de la aplicación, describiremos funcionalmente cada una de los componentes necesarios para la implementación de la solución.

6.1 Parser predictivo

La etapa del parsing consiste en recorrer la conversación, distinguir cada palabra y frase, y filtrar todos los signos de puntuación y caracteres que no aporten información adicional a la conversación para la identificación del tema.
A medida que el predictor lee la conversación entrega al parser cada unidad de información identificada, junto con sus características. Si la unidad de información es un usuario, se almacena el número de párrafo identificado. En cambio, si la unidad de información es una palabra o frase se almacena el usuario que la expresó, el párrafo en el que se dijo (asumiendo que un mismo usuario interviene repetidas veces) y la posición dentro de este párrafo.
Cada unidad de información que entrega el predictor se almacena en un objeto *Usuario* u *OcurrenciadePalabra*, según corresponda.
El objeto *Usuario* tiene los siguientes atributos:

- nombre

El objeto *OcurrenciadePalabra* tiene los siguientes atributos:

- Numero de Párrafo
- Posición dentro del Párrafo
- Nombre
- Cantidad de palabras
- Tipo (f: frase o p: palabra)
- Peso del usuario (los usuarios calificados en algún tema, sea o no el tema de la conversación que se está analizando, tendrán mayor peso que los usuarios comunes)

Ejemplo 6.1:
Se tiene el siguiente párrafo de una conversación, obtenido del libro "El Nombre de La Rosa", de Umberto Eco.

Guillermo: *La cena fue triste y silenciosa.*

El predictor compone un objeto *Usuario* u *OcurrenciadePalabra* con cada una de las unidades de información del párrafo, y lo entrega al parser.

1. Identifica la unidad de información "Guillermo", que es un usuario. Guillermo no es un usuario calificado o experto en ningún tema almacenado en la base de conocimiento, entonces todas las palabras mencionadas por este usuario tendrán el mínimo valor en el atributo "peso de usuario" (el mínimo valor es 1). Crea un objeto *Usuario* con el atributo siguiente:
   nombre: Guillermo
2. Identifica la siguiente unidad de información "La"
   Nro de párrafo: 1. Es el número de párrafo correspondiente dentro de la conversación (consideremos para este ejemplo, que es el primer párrafo de la conversación)
   Posición dentro del párrafo: 1
   Nombre: La
   Cantidad de palabras: 1
   Tipo: p (palabra)
   Peso de usuario: 1 (Guillermo no es un usuario calificado)
3. Identifica la siguiente unidad de información "cena"
   Nro de párrafo: 1 — Posición dentro del párrafo: 2
   Nombre: cena — Cantidad de palabras: 1
   Tipo: p (palabra) — Peso de usuario: 1
4. Identifica la siguiente unidad de información "fue"
   Nro de párrafo: 1 — Posición dentro del párrafo: 3
   Nombre: fue — Cantidad de palabras: 1
   Tipo: p (palabra) — Peso de usuario: 1
5. Identifica la siguiente unidad de información "triste"
   Nro de párrafo: 1 — Posición dentro del párrafo: 4
   Nombre: triste — Cantidad de palabras: 1
   Tipo: p (palabra) — Peso de usuario: 1
6. Identifica la siguiente unidad de información "y"
   Nro de párrafo: 1 — Posición dentro del párrafo: 5
   Nombre: y — Cantidad de palabras: 1
   Tipo: p (palabra) — Peso de usuario: 1
7. Identifica la siguiente unidad de información "silenciosa"
   Nro de párrafo: 1 — Posición dentro del párrafo: 6
   Nombre: silenciosa — Cantidad de palabras: 1
   Tipo: p (palabra) — Peso de usuario: 1

Por cada objeto *OcurrenciadePalabra* que el predictor entrega al parser, éste analiza si es una unidad de información para descartar o no. En caso de que no se la descarte, la unidad de información se normaliza si corresponde.

1. nombre: La
   descartar: sí
2. nombre: cena
   descartar: no
   normalización: cena

3. nombre: fue
   descartar: sí
4. nombre: triste
   descartar: no
   normalización: tristeza
5. nombre: y
   descartar: sí
6. nombre: silenciosa
   descartar: no
   normalización: silencio

Por cada objeto *Usuario* que el predictor entrega al parser, éste incrementa en 1 la cantidad total de párrafos que tiene la conversación.

El predictor es el encargado de entregar al parser las unidades de información. Entonces el predictor también será responsable de distinguir frases dentro de la conversación, dándole prioridad a las mismas por sobre las palabras que las forman.

Ejemplo 6.2:
En la base de conocimiento, existe la frase "*sistema solar*". Luego el predictor encuentra "*sistema solar*" en la conversación, y el parser la considerará como si fuera una unidad de información más. Entonces esta frase recibirá el mismo tratamiento que cualquier otra palabra de la conversación, es decir una frase será una unidad de información independiente de las palabras que la componen.

El predictor también priorizará una frase que esté compuesta por otra frase más alguna/s otra/s palabra/s. Es decir, el predictor hace un análisis especial de cada frase que encuentra, tratando de quedarse con la frase más grande que pueda a partir de la secuencia de palabras que va leyendo. Para el parser, este trabajo es transparente, y sólo recibe palabras o frases ya reconocidas en la conversación.

Ejemplo 6.3:
En la base de conocimiento se tienen las frases *"la casa"* y *"la casa encantada"*. El predictor no le devolverá al parser la frase *"la casa"* hasta no estar seguro que la siguiente palabra no sea *"encantada"*. Si en cambio aparece a continuación de *"la casa"* la palabra encantada, devolverá la frase *"la casa encantada"*.

A medida que el parser recibe del predictor cada unidad de información, lleva la cuenta de la cantidad de párrafos y palabras que hay en la conversación.
Sobre las unidades de información no descartadas, el parser calculará los factores de los patrones de comportamiento en forma incremental, usando la información previamente calculada.

6.1.1 Mecánica del Parser

El parser interactúa con el predictor, encolando cada resultado obtenido en una cola de devolución.

La forma de trabajo es e*n demanda*. Cada componente ejecuta sus responsabilidades a medida que le son requeridas.
Cuando se van incorporando unidades de información en la cola de devolución, las mismas ya están disponibles para ser almacenadas en el grafo de la conversación. Es decir en este punto, toda unidad de información almacenada se tendrá en cuenta para la representación abstracta de la conversación.

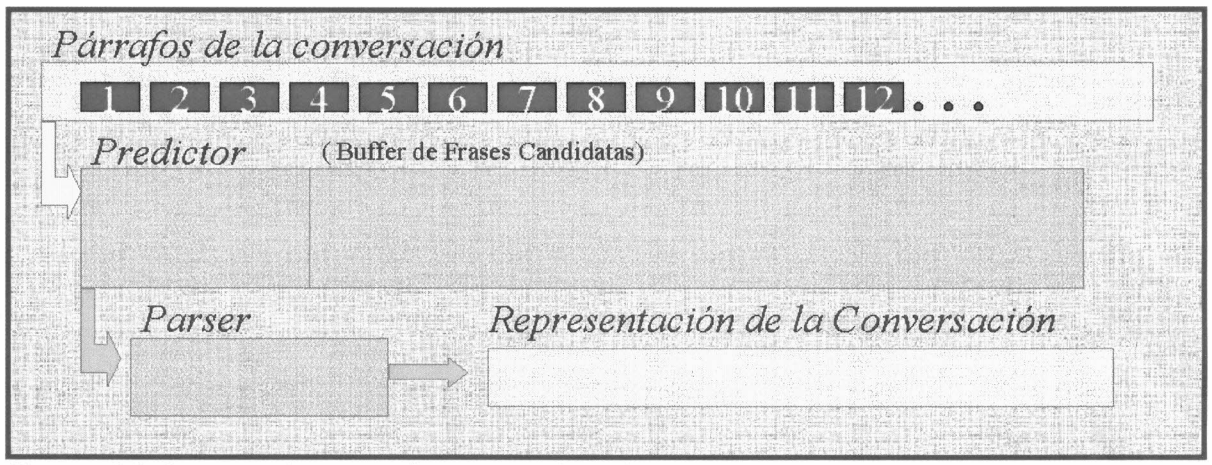


Figura 6.1. Interacción entre el parser y el predictor.
Se cuenta con una lista de palabras, que a su vez están agrupadas dentro de párrafos en la conversación. El administrador de entrada y salida devolverá cada palabra, con un identificador especial si la misma representa el nombre de un usuario.

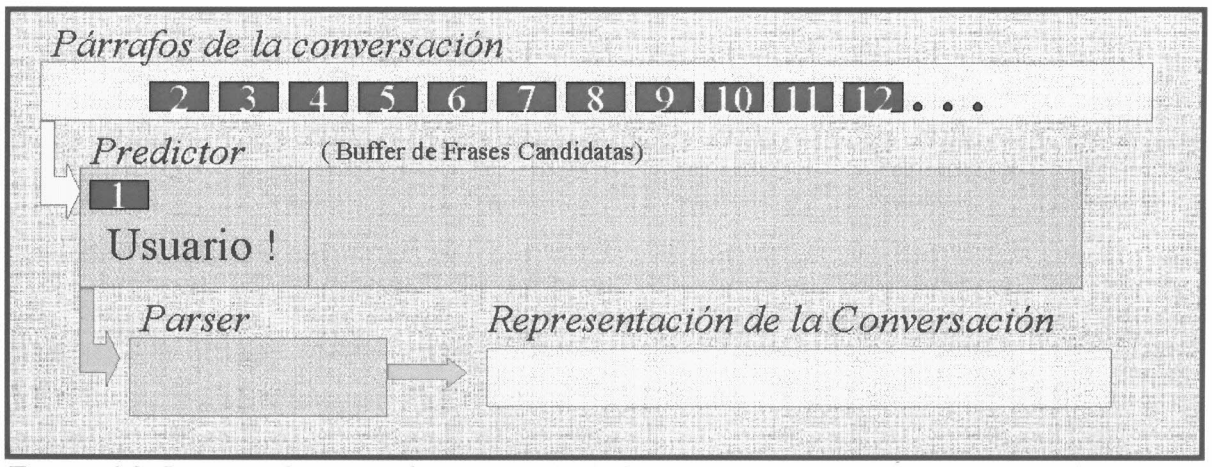


Figura 6.2. Interacción entre el parser y el predictor.
La palabra 1 es identificada por el predictor como el nombre de un usuario. De aquí en más, hasta que no se identifique un nuevo usuario, todas las palabras posteriores serán asociadas a este usuario.

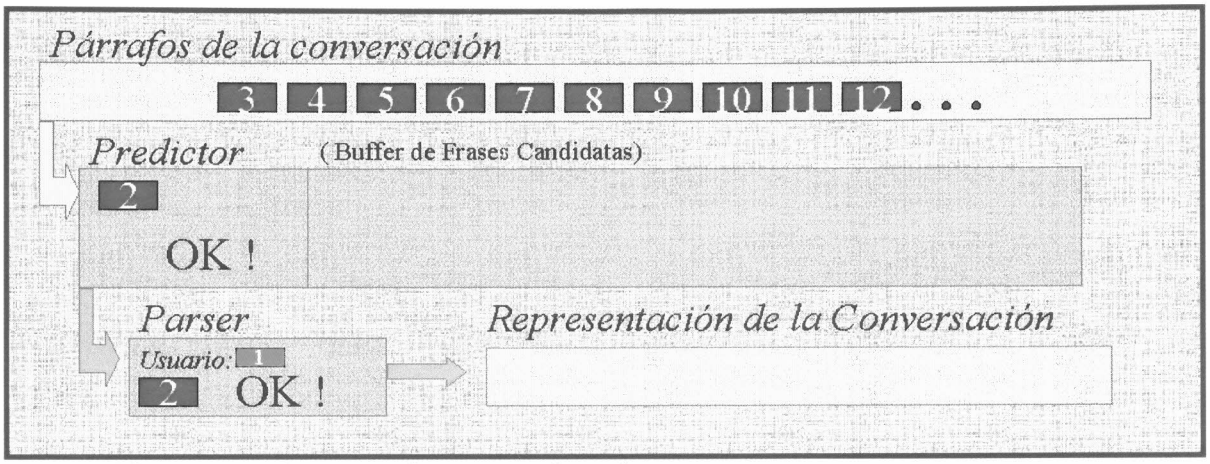


Figura 6.3. Interacción entre el parser y el predictor.
Se lee la palabra 2, que no tiene características especiales para el parser, con lo cual se deja que siga libremente hasta la representación abstracta de la conversación.

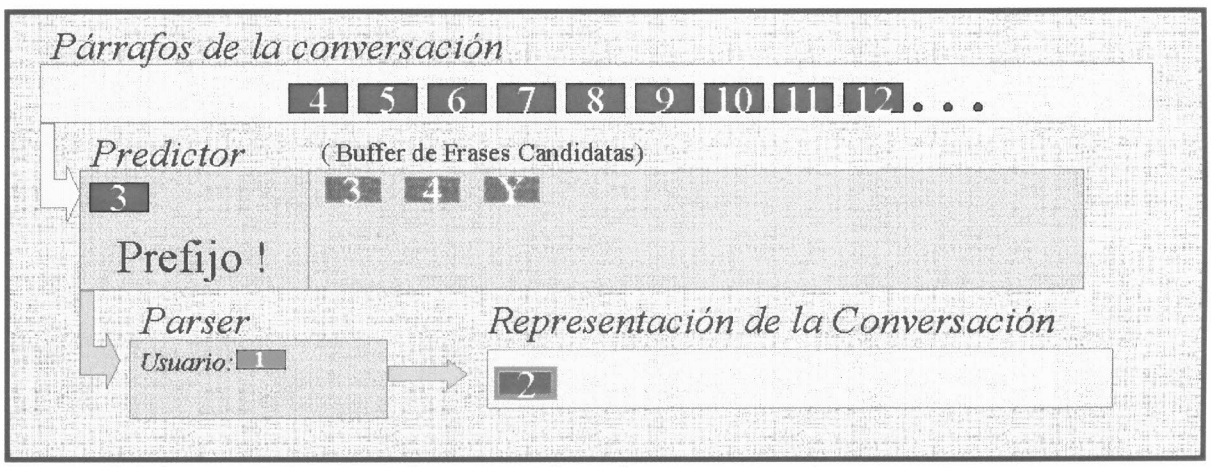


Figura 6.4. Interacción entre el parser y el predictor.
Se distingue la palabra 3 que es prefijo de alguna frase existente en la base de conocimiento. Entonces el predictor la retendrá para su análisis.

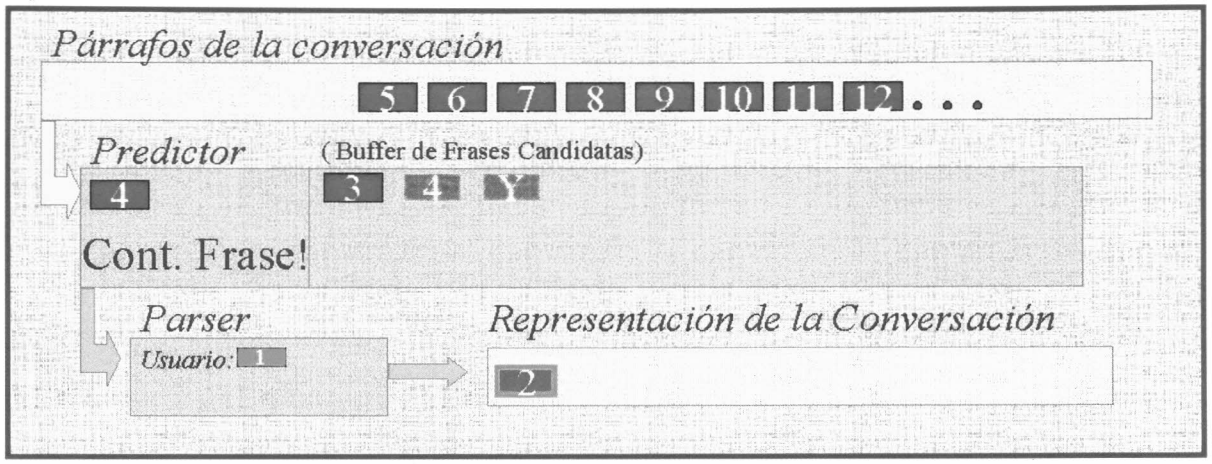


Figura 6.5. Interacción entre el parser y el predictor.
Se distingue la palabra 4 que es continuación de la frase candidata tenida en cuenta. Entonces el predictor la retiene para su análisis.

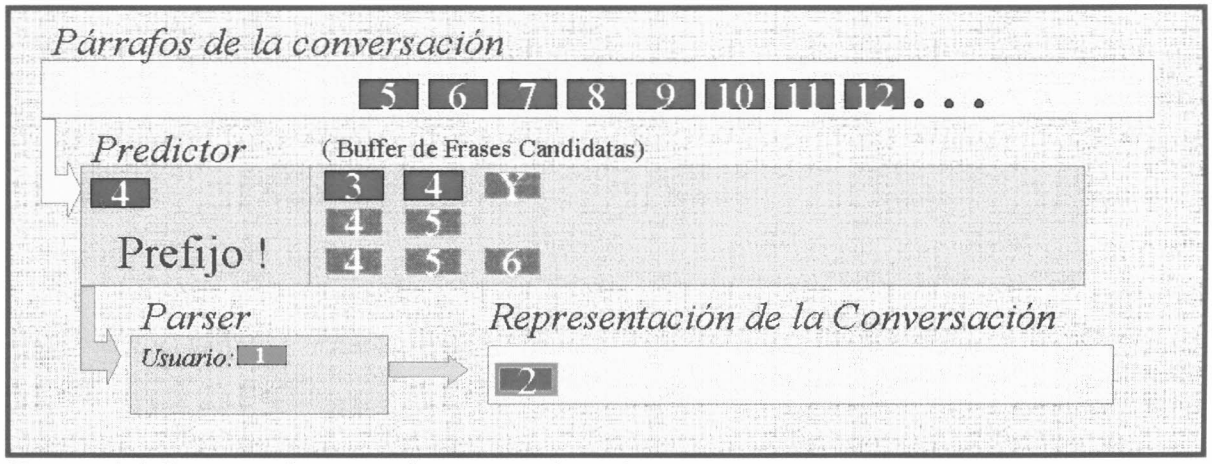


Figura 6.6. Interacción entre el parser y el predictor.
Además, esta palabra 4 es prefijo de otras dos frases existentes en la base de conocimiento, entonces por esta razón es retenida para su análisis.

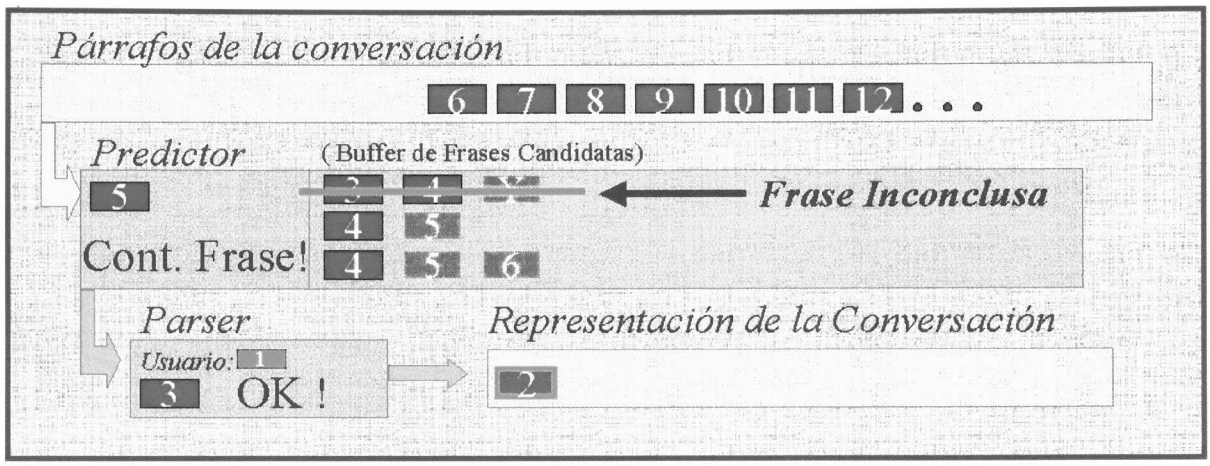


Figura 6.7. Interacción entre el parser y el predictor.

A posteriori, aparece la palabra 5. En la primer frase se esperaba la palabra Y, con lo cual esta frase deja de ser candidata. Ahora el predictor debe entregar al parser las palabras que se retenían pura y exclusivamente por esta frase descartada. En este caso es la palabra 3.

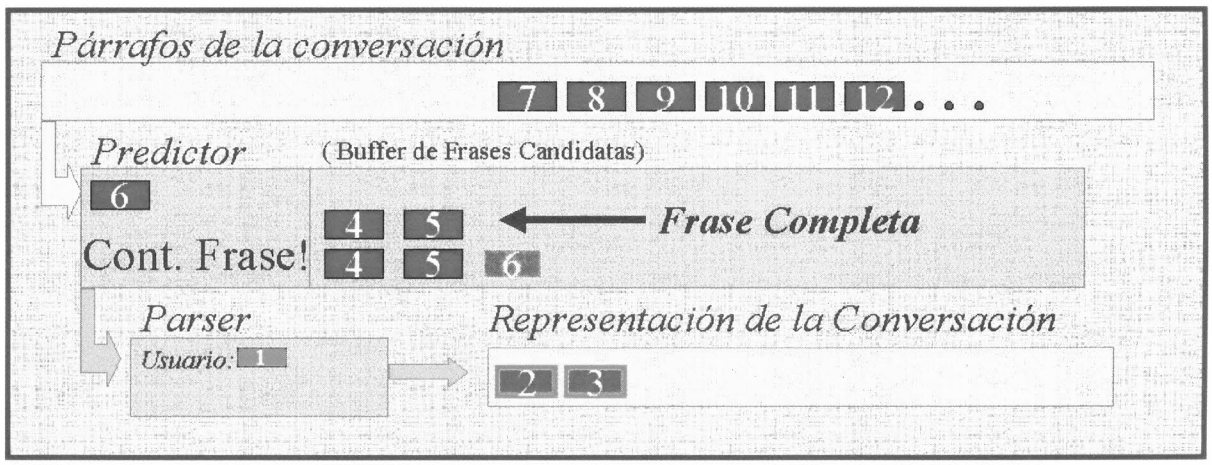


Figura 6.8. Interacción entre el parser y el predictor.

Con la aparición de esta palabra 5, se completó una de las frases candidatas. Pero al existir otra frase candidata de la cual la anterior frase es prefijo, se debe esperar, *ya que se privilegia siempre el reconocimiento de la frase más larga posible.*

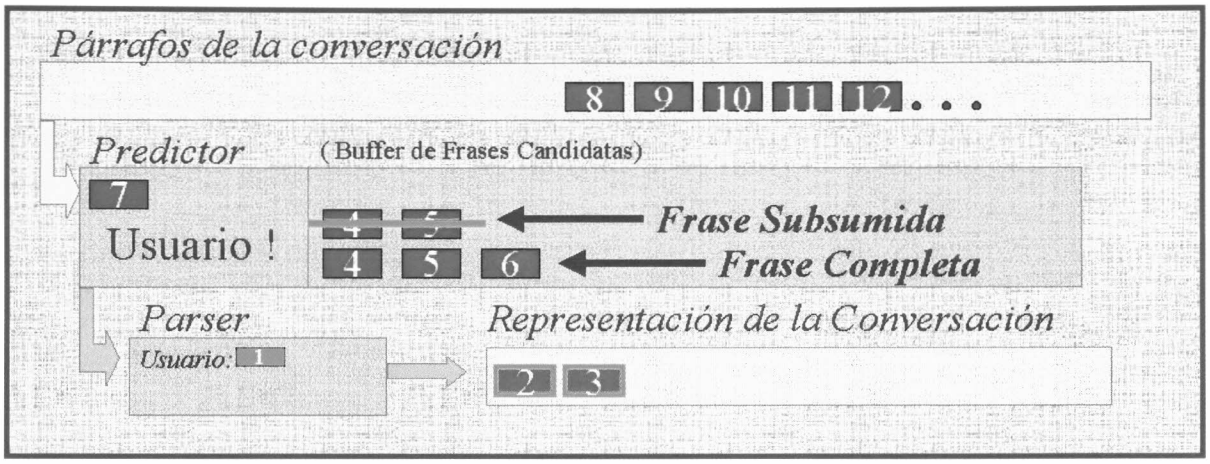


Figura 6.9. Interacción entre el parser y el predictor.

Como la palabra siguiente es la 6, se completa una frase más larga, con lo cual la anterior queda subsumida por ésta.

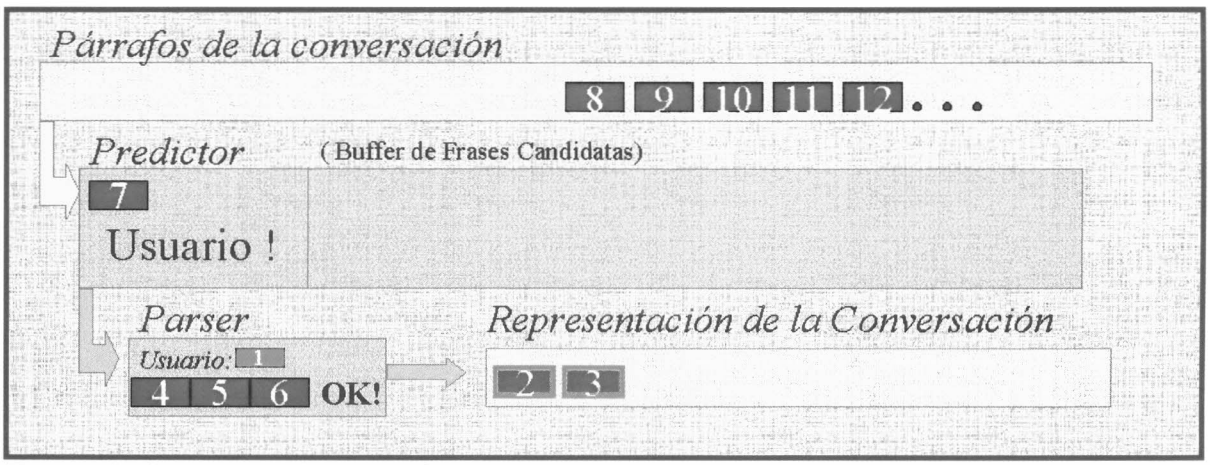


Figura 6.10. Interacción entre el parser y el predictor.

La palabra 7 es un nuevo usuario. Por otra parte, el predictor entrega al parser la frase reconocida de la misma manera que entregaría una palabra común. Para el parser esta distinción es transparente.

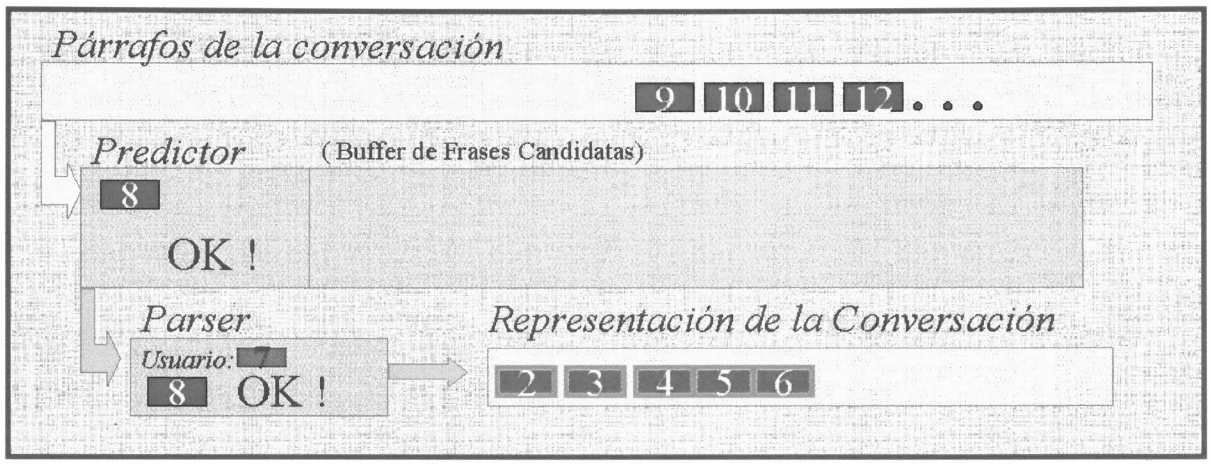


Figura 6.11. Interacción entre el parser y el predictor.

A continuación se lee la palabra 8, que no tiene características especiales para el predictor ni para el parser, así que se almacena directamente en la representación abstracta de la conversación.

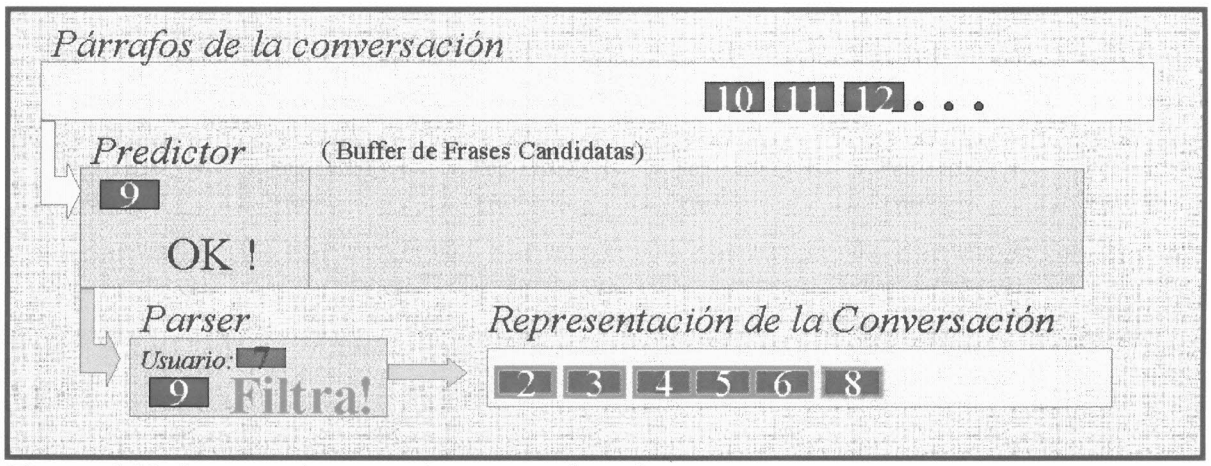


Figura 6.12. Interacción entre el parser y el predictor.

Se lee la palabra 9, que no tiene características especiales para el predictor, pero el parser detecta que es una palabra que debe ser descartada.

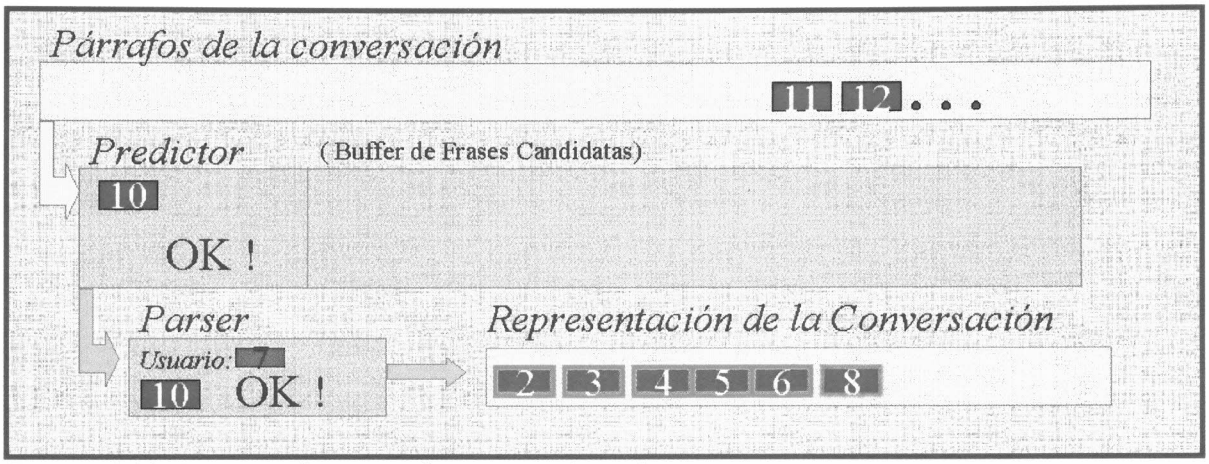


Figura 6.13. Interacción entre el parser y el predictor.

A continuación se lee la palabra 10, que no tiene características especiales para el predictor ni para el parser. Entonces se almacena directamente en la representación abstracta de la conversación.

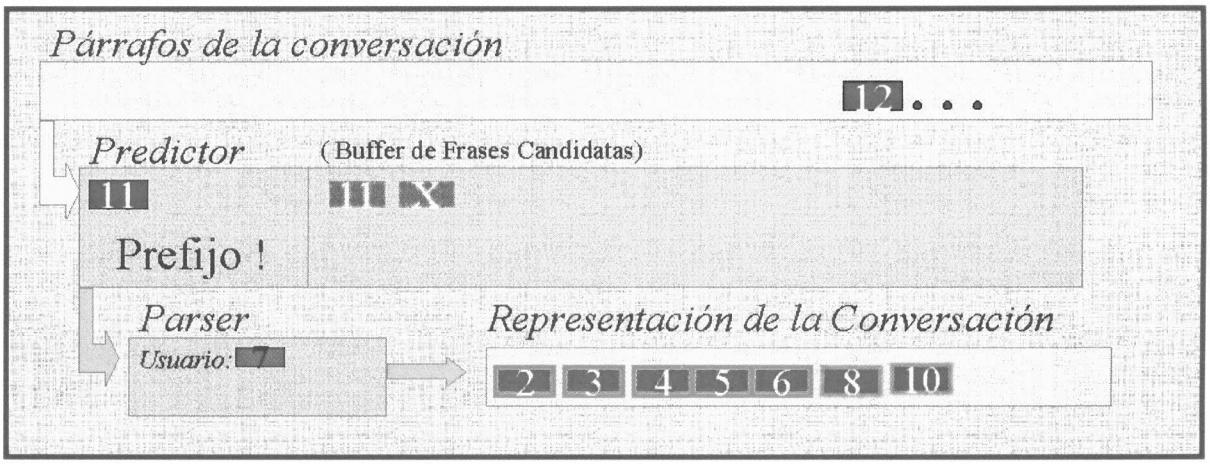


Figura 6.14. Interacción entre el parser y el predictor.

Se lee la palabra 11, que es prefijo de alguna frase existente en la base de conocimiento, entonces el predictor la retendrá para su análisis.

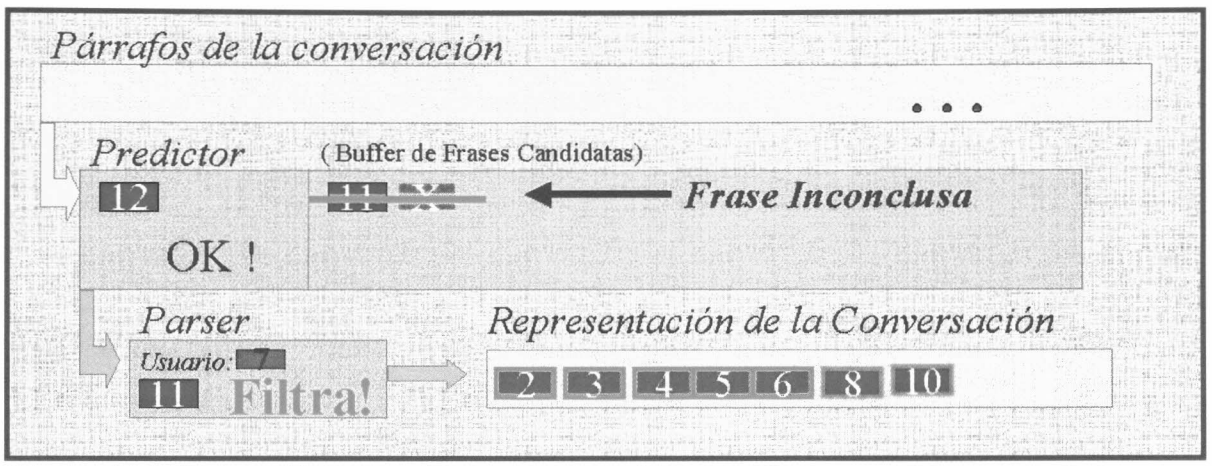


Figura 6.15. Interacción entre el parser y el predictor.

A continuación aparece la palabra 12. En la frase candidata se esperaba la palabra X, con lo cual esta frase deja de ser candidata.

Ahora el predictor debe entregar al parser las palabras retenidas por esta frase descartada, en este caso la palabra 11, y el parser detecta que esta palabra es para descartar, con lo cual no la agrega en la representación abstracta de la conversación.

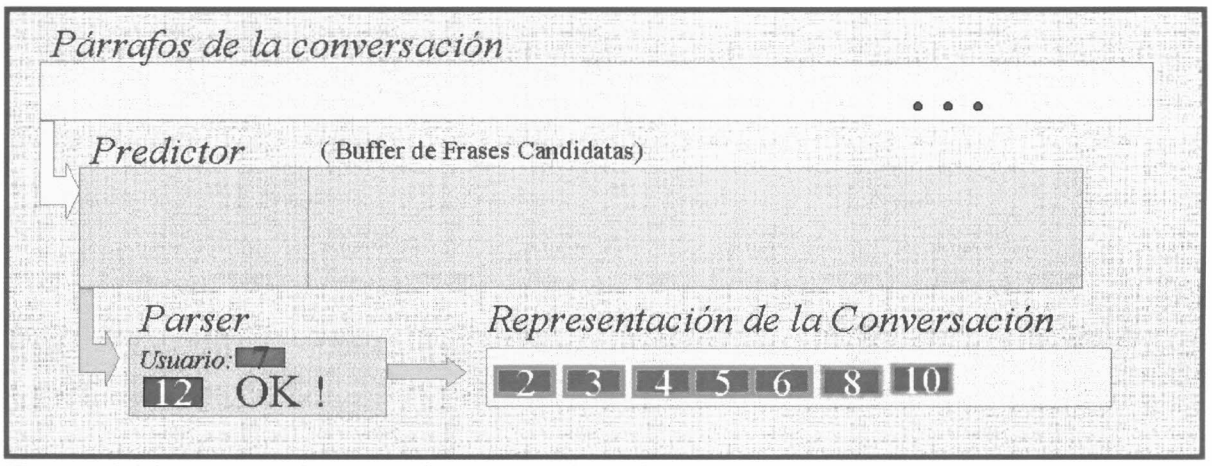


Figura 6.16. Interacción entre el parser y el predictor.

A continuación el predictor entrega al parser la palabra pendiente, palabra 12.

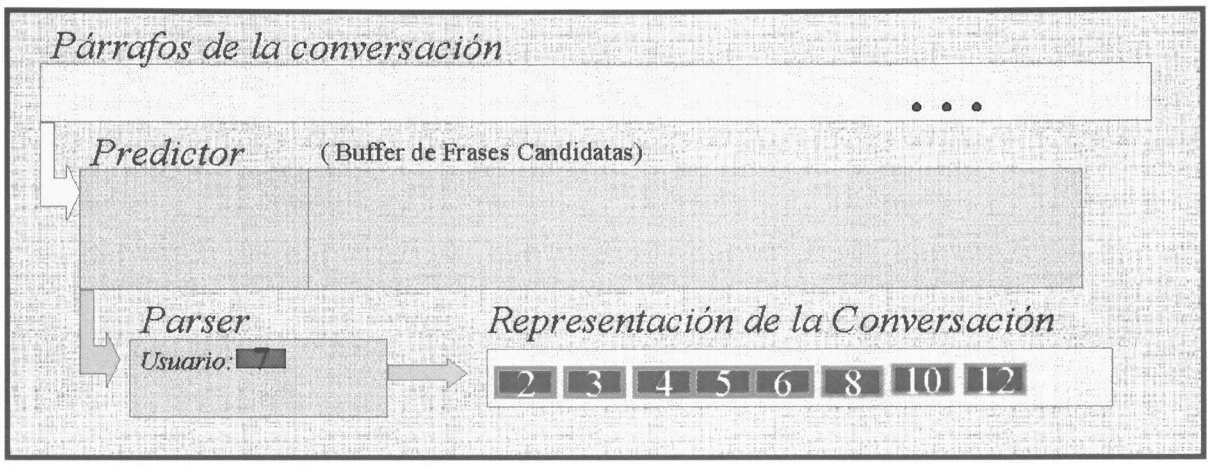


Figura 6.17. Interacción entre el parser y el predictor.

En la representación abstracta de la conversación quedaron todas las unidades de información a partir de las cuales se hará la interpretación de la conversación.

6.2 Interpretación

Para aliviar el trabajo de la interpretación, las palabras que tengan una frecuencia promedio menor a una frecuencia mínima preestablecida se descartarán, exceptuando aquellas que tengan el atributo de clave debido a su potencial importancia en la búsqueda de temas.
En la interpretación se asigna un peso a las palabras de tal forma que un mayor peso represente una mayor importancia de esa palabra, y así poder seleccionar un subconjunto relevante del total de palabras existentes en la conversación, quedando formado este subconjunto por las palabras de mayor peso, y que las mismas representen semánticamente a la conversación.

6.2.1 Cálculo de los pesos según cada patrón

El peso resultante para cada palabra o frase es el producto de los pesos asociados a cada uno de los patrones considerados (ver patrones en Sección 4.1.2).
Los pesos de las palabras se usarán para podar el grafo. Las palabras que sobrevivan serán las que se usarán para comparar contra las palabras que identifican a cada uno de los temas existentes en la base de conocimiento.

Definición 6.1: [Peso neutro de una palabra]
Peso neutro de una palabra es un valor que indica que esa palabra no tiene un significado favorable ni desfavorable en la conversación.

Se establece arbitrariamente el peso neutro de una palabra en 1, entonces cada uno de los factores que influyen en el peso de las palabras deberá ser un valor mayor o igual a 1 (ya que al peso inicial de cada palabra, que es 1, se lo multiplica por la incidencia asociada a cada uno de estos factores).
Entonces, es necesario que cada factor se exprese por un valor normalizado que tenga 1 como cota inferior. Como no se tiene restricción alguna para la cota superior, la establecemos arbitrariamente.
La idea de los factores es que produzcan un corrimiento moderado sobre los pesos de las palabras. De esta forma tendrán una mayor incidencia en la determinación del tema al tener un valor superior a 1. Pero debemos cuidar que estos valores resultantes no produzcan un corrimiento exagerado, porque sino tendrían una ventaja muy abrupta respecto de las palabras cuyo peso se mantuvo en 1, pero que igualmente puedan conducir al tema correcto.

6.2.1.1 Cálculo de distancia promedio entre par de palabras o frases

El cálculo de la *Distancia Promedio entre cada par de palabras o frases* (def. 4.4) está dada por la siguiente fórmula, donde n es la cantidad de ocurrencias de la palabra a, m es la cantidad de ocurrencias de la palabra b, a_i es el número de párrafo donde se encuentra la i-ésima ocurrencia de la palabra a, y b_j es el número de párrafo donde se encuentra la j-ésima ocurrencia de la palabra b.

$$\left(\sum_{i=1,n} \sum_{j=1,m} |a_i - b_j|\right) / n * m \qquad (1)$$

Como las palabras se van analizando secuencialmente, se necesita optimizar esta expresión para recalcular la distancia promedio de cada palabra a partir de una ocurrencia nueva y de los cálculos ya realizados. A continuación se verá la forma de realizarlo de manera incremental.

6.2.1.1.1 Cálculo incremental de la distancia promedio

Se recalculan todas las distancias y promedios al agregarse una nueva ocurrencia de una palabra, usando la información previamente calculada, evitando así recalcular sobre la base de toda la información existente.

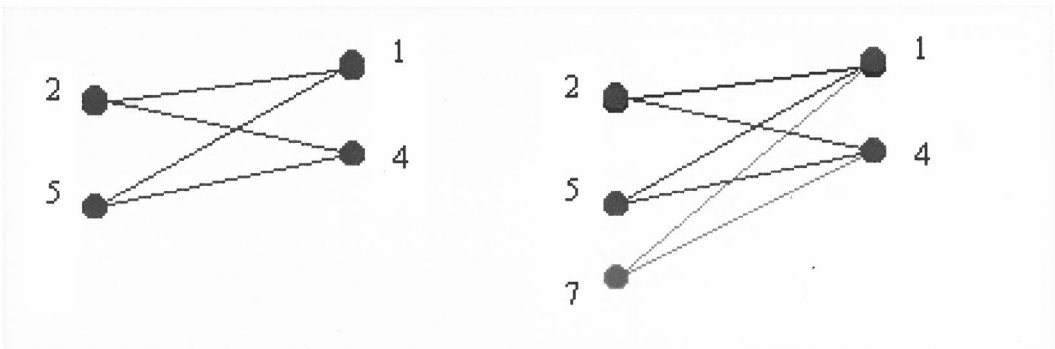


Figura 6.18. Cálculo incremental de la distancia promedio

Cada nodo en el gráfico representa la ocurrencia de una palabra. A la izquierda las de la palabra *A*, y a la derecha las de la palabra *B*. Cada nodo está rotulado con su posición (el número de párrafo al que pertenece esa ocurrencia de palabra). Cada arco representa la distancia promedio de las ocurrencias de palabras representadas por los nodos unidos por este arco.

Las variables que se usan son:
A.promedio: es el promedio de todas las posiciones de la palabra *A*.
B.promedio: es el promedio de todas las posiciones de la palabra *B*.
Estadistica.promedioDistancia(A, B): es promedio de las distancias entre todas las posiciones de la palabra *A* y todas las posiciones de la palabra *B*.
A.cantidad: es la cantidad de ocurrencias de la palabra *A*.
B.cantidad: es la cantidad de ocurrencias de la palabra *B*.

Ejemplo 6.4:
Ana: ¿La receta de torta frita lleva levadura? (párrafo 1)
Mirta: La receta que uso yo no. (párrafo 2)

A = "receta"
B = "torta frita"

A.promedio = (1+2) / 2 = 1.5
B.promedio = 1/1 = 1
Estadistica.promedioDistancia(A, B) = (|1 - 1| + |2 - 1|) / (1*2) = 0.5
A.cantidad = 2
B.cantidad = 1

El algoritmo para calcular el promedio de distancia entre 2 palabras, es inductivo. Calcula la distancia promedio de la palabra actual contra todas las demás palabras a partir del cambio producido por la nueva ocurrencia. Es decir, se calcula la distancia promedio entre Pm y Pi (i<>m). Hay que tener en cuenta que la posición de la nueva palabra que se está analizando (Pm,n), es mayor que la posición de cualquier otra palabra que ya ha sido analizada anteriormente.

```
Mientras Conversacion <> vacio hacer
    Ingresa una nueva palabra P_m,n
    //donde m es el nro. de palabra y n el nro. de ocurrencia de la misma
    P_m.cantidad= n

    Si n = 1 entonces
        P_m.promedio = P_m,n.posicion
    Sino
        P_m.promedio = (P_m.promedio * (n-1)) + P_m,n.posicion) / n
    Fin si

    Para i= 1 hasta Conversacion.cantidad_palabras, i <> m
        Si n = 1 entonces
            sumaDistancias = 0
            Para j = 1 hasta P_i.cantidad
                sumaDistancias = P_m,1.posicion – P_i,j.posicion
            Fin para
            Estadistica.promedioDistancia(P_i,P_m) = sumaDistancia / P_i.cantidad
        sino
            // Fórmula de distancia promedio
            Estadistica.promedioDistancia(P_i,P_m) =

                ( ( Estadistica.promedioDistancia(P_i,P_m) * (n – 1) * P_i.cantidad)
          + ((P_m,n.Posicion – P_i.promedio) * P_i.cantidad) )
            ---------------------------------------------------------------------
                                    n * P_i.cantidad
        Fin si
    fin Para
    Siguiente i
Fin mientras
```

Demostración

El primer término en la suma del numerador es la suma de las distancias de los arcos entre las ocurrencias de P_i, y las $(n-1)$ ocurrencias de P_m.

Estadistica.promedioDistancia(P_i,P_m) * (n – 1) * P_i.cantidad

que representa la suma de las distancias de los arcos entre las ocurrencias de P_i, y las (n-1) ocurrencias de P_m entonces,

$$\text{Estadistica.promedioDistancia}(P_i,P_m) = \sum d(P_i, P_m) \;/\; \left((n-1) * P_i.\text{cantidad} \right)$$

donde $d(P_i, P_m)$ es la distancia de los arcos entre las ocurrencias de P_i y las (n-1) ocurrencias de P_m.

El segundo término en la suma del numerador es la suma de las distancias de los arcos que se agregaron entre las ocurrencias de P_i, y la nueva ocurrencia de $P_{m,n}$.

$$(P_{m,n}.Posicion - P_i.promedio) * P_i.cantidad$$

porque la suma de las distancias de los arcos que se agregaron entre las ocurrencias de P_i, y la nueva ocurrencia $P_{m,n}$ es

$$\sum_{j=1,Pi.cantidad} (P_{m,n}.Posicion - P_{i,j}.Posicion)$$

y esto es igual a:

$$P_{m,n}.Posicion * Pi.cantidad - \sum_{j=1,Pi.cantidad} (P_{i,j}.Posicion)$$

$$(P_{m,n}.posicion - \sum_{j=1,Pi.cantidad} (P_{i,j}.posicion) / P_i.cantidad) * P_i.cantidad$$

y esto es igual a

$$(P_{m,n}.posicion - P_i.promedio) * P_i.cantidad$$

Finalmente, el denominador es la cantidad de arcos entre P_m y P_i, es decir, la cantidad de ocurrencias de P_m (= n) por la cantidad de ocurrencias de P_i (= P_i.cantidad)

$$n * P_i.cantidad$$

6.2.1.2 Frecuencia

Este cálculo se realiza para cada palabra o frase, una vez concluida la lectura de la conversación, que es cuando se conoce el total de ocurrencias de palabras en la conversación.
Así, la frecuencia asociada a una palabra *a*, la podemos expresar como:

$$frecuencia_a = Cant.Ocurr_a / Cant.tot.OcurrPalabras$$

6.2.1.3 Concentración

La fórmula de la concentración de una palabra (ver sección 4.1.2.3) es:

$$\left(\sum_{i=1,n-1} |a_{i+1} - a_i|\right) / ((n-1) * n)$$

donde *n* es la cantidad de ocurrencias de la palabra *a*, *(n-1)* es la cantidad de distancias calculadas, a_i es el número de párrafo donde se encuentra la *i-ésima* ocurrencia de la

palabra a, y a_{i+1} es el número de párrafo donde se encuentra la *(i+1)-ésima* ocurrencia de la palabra a, y $|a_{i+1} - a_i|$ es la distancia entre la aparición de una ocurrencia y la siguiente.

6.2.1.4 Condición de Palabra Clave

Este valor no requiere de ningún cálculo previo, porque está asociado a cada uno de los temas de los cuales es clave la palabra o frase. Pero el atributo de clave servirá para mantener dicha palabra o frase dentro de la interpretación.

6.3 Búsqueda de temas

De las palabras resultantes de la interpretación, se eligen las *palabras de búsqueda del tema*.

Definición 6.3: [Palabras de búsqueda del tema]
Las *palabras de búsqueda del tema* serán las palabras destacadas de la interpretación elegidas de acuerdo a algún criterio, y ordenadas de acuerdo al peso dado por la interpretación.

La idea es que las palabras de búsqueda no tienen que ser todas las palabras destacadas de la conversación, ya que puede haber limitaciones en la cantidad de palabras que pueda soportar el proceso de búsqueda. Además, se debe diferenciar el sentido dado al nombrar unas y otras. Cuando se habla de palabras destacadas, se habla del resultado del proceso de interpretación hecha por la aplicación. En cambio al hablar de palabras de búsqueda, se hace referencia en las palabras que se usarán para comparar contra los temas existentes en la base de conocimiento. Para nuestra aplicación, todas las palabras destacadas son usadas en el proceso de búsqueda.
Con las mismas se procede a buscar en la base de conocimiento el tema que tenga mayor puntaje, teniendo en cuenta el peso de las palabras asociadas a cada tema almacenado.
Se analizan dos métodos: uno basado en la búsqueda del tema que tenga la mayor cantidad de palabras coincidentes. Este método es lento debido al carácter exhaustivo de la búsqueda. El segundo método, finalmente elegido, delega la responsabilidad a la base de datos usada para almacenar la base de conocimiento, y se comporta de manera más eficiente. Ambos métodos se explican a continuación.

6.3.1 Método exhaustivo

A partir de las palabras resultantes de la interpretación, se procede a buscar en la base de datos todos los temas que tengan el total de las palabras. Se repite el proceso con el total menos 1 (todas las combinaciones posibles) y así sucesivamente, hasta que se encuentre uno o más temas con una cantidad i de palabras, para todas las combinaciones en que se hayan encontrado temas. En este punto se detendrá la búsqueda, y, de todos los temas encontrados con i palabras, se elegirá aquel cuya medida de proximidad sea la mejor de todas. Este método de búsqueda es exhaustivo, y como consecuencia extremadamente lento.

6.3.2 Método basado en Lenguaje de Consultas SQL

En este caso se realiza una consulta en la base de datos conteniendo la lógica descripta en el método exhaustivo y obteniendo, si existe,el tema de mayor puntaje con las palabras dadas. La eficiencia se delega en este caso al motor de la base de datos.

6.4 Decisión

En caso de haber tema candidato, se determinará la relación que guarda con la conversación analizada, según las relaciones de los faltantes, sobrantes y coincidentes con los intervalos de aceptación, incertidumbre y rechazo, tal como se explicó en el diseño de la aplicación. Si no hay tema candidato, se analiza entonces la cantidad de información proveniente de la conversación para saber si la misma es suficiente o no para determinar un tema. En caso de ser suficiente, se le preguntará al usuario calificado el nombre del tema. La información de las decisiones tomadas por el usuario se guardan en la base de conocimiento.

6.5 Autoaprendizaje

La aplicación permitirá la adquisición de nuevo conocimiento. Se busca que este nuevo conocimiento sea consistente con el ya existente en la base de conocimiento. A continuación se describirán las distintas situaciones que se pueden presentar.

6.5.1 Inclusión de nuevo vocabulario

A medida que se recorre la conversación, se guarda en la base de conocimiento las palabras encontradas que sean desconocidas, para una posterior identificación de las mismas.

6.5.2 Inclusión de nuevos temas

Si la conversación está relacionada a un tema no existente en la base de conocimiento, y el usuario calificado informa el nombre del mismo, se incorporará el nuevo tema a la red semántica de la base de conocimiento.

6.5.3 Inclusión de nuevas relaciones entre los temas

Si la conversación está relacionada a un tema no existente en la base de conocimiento, pero asociada a un tema ya existente (por ejemplo una relación de tema más específico o tema más general) y el usuario calificado informa el nombre del mismo, también se incorporará, además del nuevo tema, la relación del mismo con el tema relacionado en la red semántica de la base de conocimiento.

6.5.4 Experiencia a partir de las decisiones del usuario

Los casos de incertidumbre pueden usarse para estudiar cuales son los márgenes de aprobación y rechazo por parte del usuario, y usar esta información para efectuar correcciones sobre las cotas preestablecidas. De esta forma se espera que la aplicación

aumente su similitud con el comportamiento que tiene el usuario calificado, acumulando la experiencia.
Gracias al autoaprendizaje, la base de conocimiento no sólo se alimenta con la información explícita dada por el usuario en la incorporación de nuevos temas, sino que además incorpora información inferida a partir de las decisiones tomadas por el usuario.
La consideración de la inferencia estadística en la base de conocimiento asegura la perdurabilidad en el tiempo, de la aplicación. De otra forma, la herramienta sólo es útil en un contexto estático de información.

6.6 Resumen

En el presente capítulo se presentó la funcionalidad de cada una de las componentes de la aplicación. También se explicó las distintas situaciones que permiten a la aplicación adquirir nuevo conocimiento.
En el próximo capítulo se explicarán los detalles de la construcción de la herramienta: descripción de las clases principales y estructura de la base de datos.

Capítulo 7

Construcción de la Herramienta

7.1 Introducción

La aplicación fue desarrollada en lenguaje Java y para almacenar y recuperar la información se utilizó una base de datos relacional, en este caso MS-SQL Server 7.0. Se eligió éste lenguaje porque cumple con los requisitos de ser orientado a objetos, lo que facilita un buen encapsulamiento de cada uno de los componentes de la aplicación, permitiendo además que los mismos puedan ejecutarse en forma distribuída.

7.2 Descripción de las Clases Principales

Simulador se encarga de proveer una interfase con el usuario, inicializar los componentes principales de la herramienta, invocar a la aplicación e informar los resultados.
Aplicacion encapsula la funcionalidad de la aplicación, ofreciendo a la clase Simulador, un punto de entrada único.
AdminES centraliza y encapsula el acceso de la aplicación a la base de conocimiento (almacenada en la base de datos), a los archivos de las conversaciones y a los archivos de resultados.
BaseDeDatos proporciona una interfase de acceso a la base de datos utilizada, siendo invocada en forma exclusiva por la clase AdminES.
ArchivoTexto brinda una interfase para el manejo de archivos de texto, siendo invocada en forma exclusiva por la clase AdminES.
Conversacion contiene toda la información de la conversación que se analiza.
Parametros contiene toda la información de los parámetros seleccionados para la corrida actual de la aplicación.
Parser, *Interpretacion*, *Buscador* y *Decision* se encargan de cumplir con todo el proceso funcional definido para el componente respectivo de la aplicación (ver capítulos 3, 4 y 5).
Resultado almacena las decisiones y resultados obtenidos en la corrida de la aplicación.

7.3 Diagrama de Colaboración de las Clases Principales

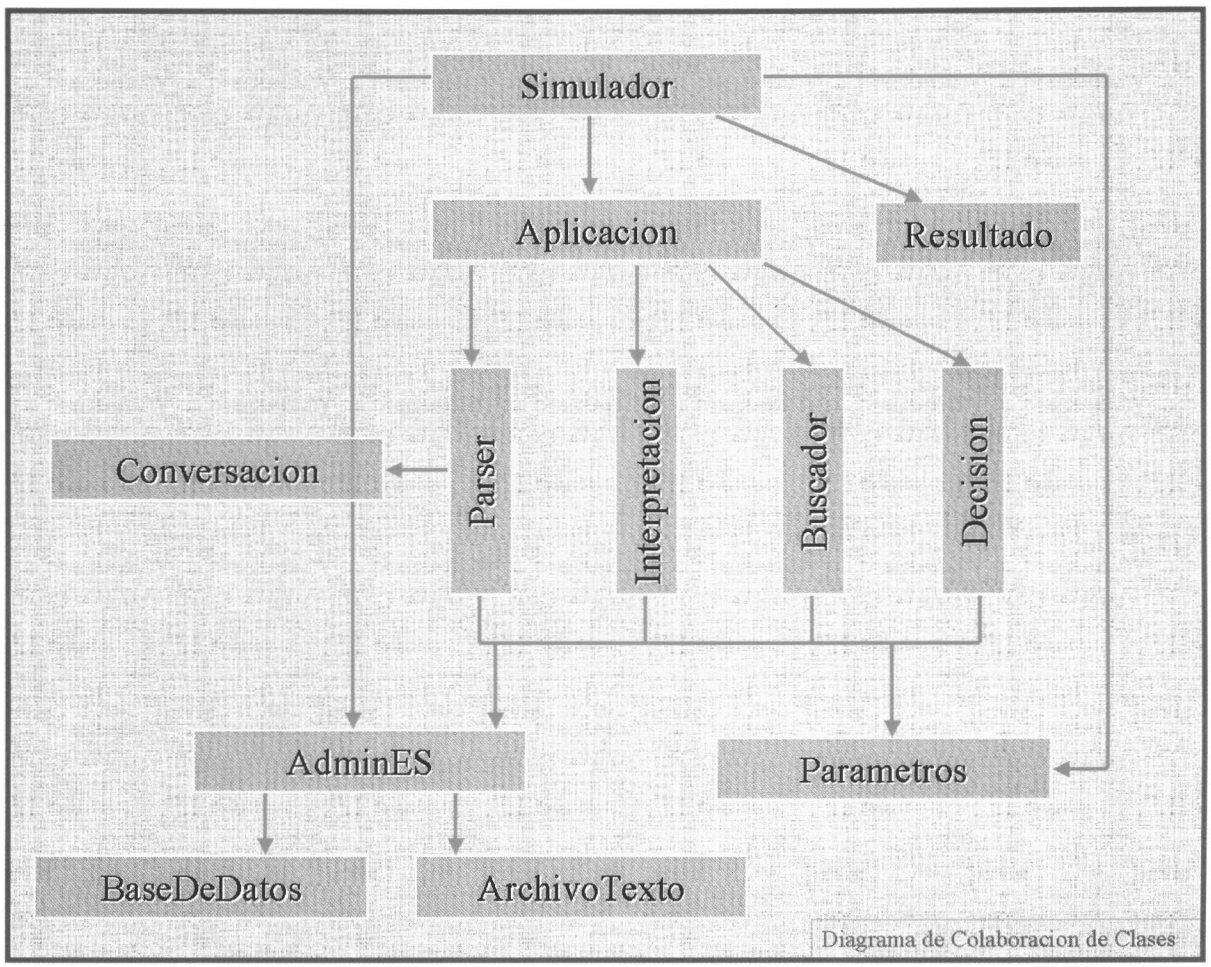


Figura 7.1. Diagrama de colaboración de las clases principales

7.4 Estructura de la Base de Datos

7.4.1 Temas

El identificador de un tema y su nombre se almacenan en la tabla *Tesis_temas*.
La jerarquía de temas está almacenada en la tabla *Tesis_relaciones*.
Tiene un identificador de tema padre y un identificador de tema hijo, permitiendo de esta forma que un tema pueda tener varios temas padres y/o varios temas hijos.

7.4.2 Palabras de temas

Cada tema tiene un conjunto de palabras que lo identifican. Estas palabras están en la tabla *Tesis_PalabrasdeTemas*. La idea es que un tema con jerarquía mayor tenga un conjunto de

palabras identificadoras que incluyan al conjunto de palabras identificadoras del padre. Es decir, al ser el tema hijo más específico que el tema padre, habrá de tener todas las palabras del padre, más algunas otras.

7.4.3 Diccionario Normalizador

Existe una tabla *Tesis_DiccionarioNormalizador* que es un diccionario. Se almacenan todas las palabras que puede aceptar la herramienta, y su palabra raíz.
En esta tabla también estarán los sinónimos. Los sinónimos formarán clases de equivalencia. Todos los sinónimos de una palabra formarán una clase de equivalencia, y tendrá un "representante", que es el que estará en todas las tablas.
Entonces, en esta tabla se almacenan las palabras (sin normalizar), y su sinónimo "representante" (normalizado).

Ejemplo 7.1:
Si una clase de equivalencia está formada por las palabras: "lindo", "hermoso", "bello", "precioso", y se asigna como palabra representante de esta clase la palabra "lindo", entonces en el diccionario normalizador habrá 4 registros, uno para cada palabra, y todos tendrán en el campo *representante* a "lindo".

Las frases también se encuentran normalizadas en la tabla *Tesis_ClavesyFrases*.
Las palabras que se almacenan en las distintas tablas de la base de datos están normalizadas.

7.4.4 Usuarios

La tabla *Tesis_Usuarios* almacena a usuarios expertos en determinados temas, los temas en que son expertos y el grado de experiencia que tienen en cada uno de los temas.

7.4.5 Jerarquía de temas

La jerarquía de temas está representada en la red semántica que se representa con las tablas *Tesis_temas*, *Tesis_PalabrasdeTemas* y *Tesis_Relaciones*.
Cada tema tiene asociado un conjunto de palabras que lo identifican, y "claves" con determinado peso.

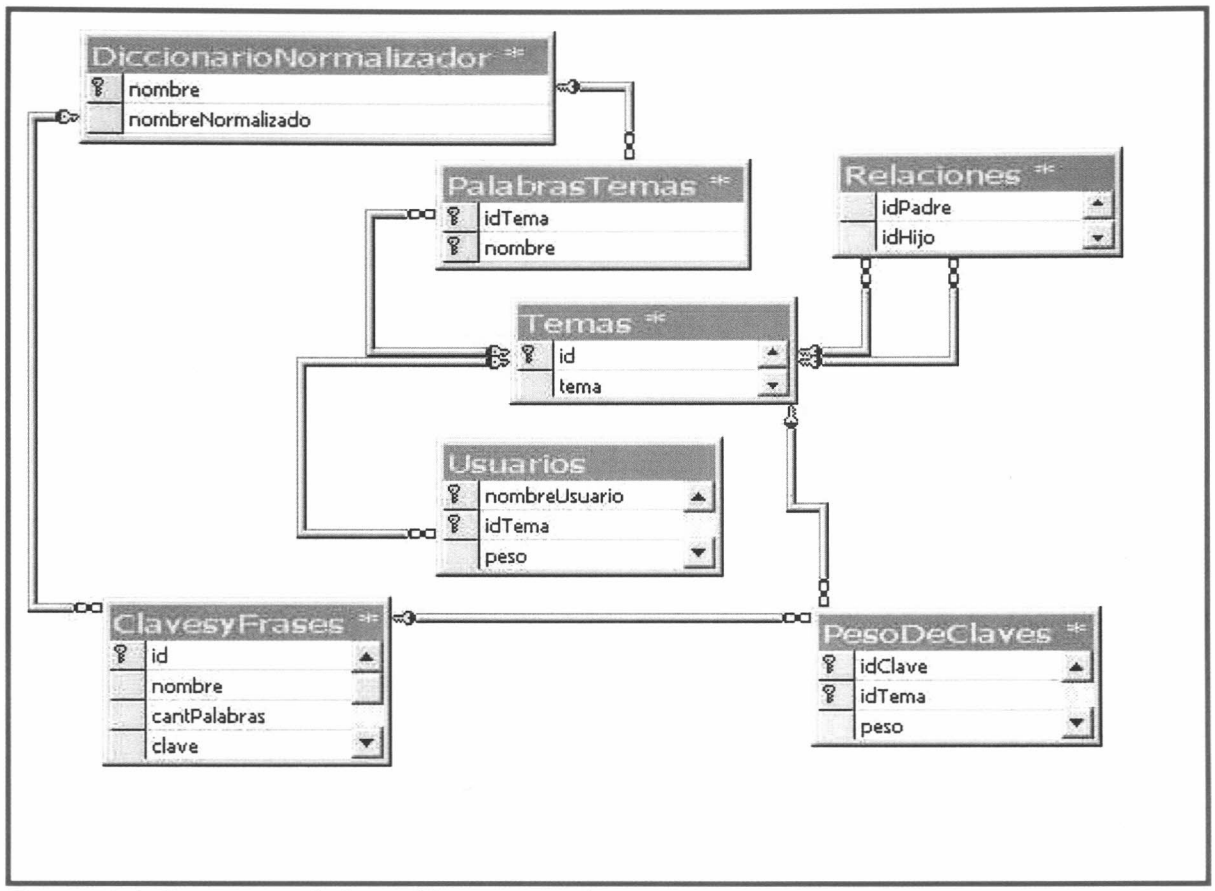


Figura 7.2. Diagrama de las tablas principales

7.5 Resumen

La herramienta se construyó utilizando diseño orientado a objetos, tratando lograr una correcta distinción de cada uno de los componentes de la aplicación y su respectivo encapsulamiento. De esta forma, si se quiere reemplazar un componente por otro que contemple una nueva estrategia o esté optimizado en el uso de recursos, se podrá hacer siempre y cuando respete el mismo protocolo que el componente original. Además, la independencia entre los distintos componentes permiten que se pueda ejecutarse en forma distribuida.

Capítulo 8

Experimentación

La herramienta fue probada y afinada con lotes de prueba compuestos por 40 conversaciones que pertenecen a 10 temas distintos, habiendo entre 2 y 6 conversaciones por tema, y obtenidos de manera aleatoria de los foros de discusión de un diario de primera línea por intenet.
Cada caso de prueba consiste en la corrida de todas las conversaciones a partir de una base de conocimiento vacía, y un conjunto de palabras y frases para filtrar establecido de manera arbitraria, que se mantuvo inalterable en las distintas pruebas.
A medida que se corrieron los distintos casos de prueba, se analiza el comportamiento de la aplicación a partir de los resultados obtenidos, y se estudiaron las modificaciones necesarias en los parámetros para afinar la herramienta a efectos de obtener una mejora en los resultados de las pruebas subsiguientes.

8.1 Variables analizadas

Para cada corrida se calcularon los valores de las siguientes variables.

Temas desconocidos
Es la cantidad de conversaciones para las cuales la aplicación no pudo encontrar un tema relacionado.

Temas no encontrados
Es la cantidad de conversaciones para las cuales la aplicación no encontró un tema relacionado y debería haberlo hecho.

Temas encontrados
Es la cantidad de conversaciones cuyos temas fueron encontrados en forma certera.

Temas encontrados con incertidumbre
Es la cantidad de conversaciones cuyos temas fueron encontrados de manera no certera.

Temas encontrados mal con incertidumbre

Cantidad de conversaciones que de manera no certera se relacionaron a temas que no correspondían.

Temas encontrados mal con certeza

Cantidad de conversaciones que se relacionaron a un tema no correspondiente de manera certera.

Temas relacionados

Representa la cantidad de conversaciones que fueron *bien* relacionadas como pertenecientes a un tema padre o un tema hijo de una conversación
existente en la base de conocimiento.

Temas mal relacionados

Representa la cantidad de conversaciones que fueron *mal* relacionadas como pertenecientes a un tema padre o un tema hijo de una conversación existente en la base de conocimiento.

8.2 Detalle de las pruebas

A continuación se describen las pruebas realizadas con un total de 520 análisis realizados por la herramienta sobre las 40 conversaciones. En la tabla 8.1. se describen los parámetros usados para cada corrida.

parámetro / nro. de corrida:	1	2	3	4	5	6	7	8	9	10	11	12	13
maxPalsBusqueda:	25.0	25.0	16.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0	35.0
cotaSupDescarteSobrante:	25.0	25.0	25.0	25.0	35.0	35.0	35.0	35.0	50.0	65.0	65.0	65.0	65.0
frecuenciaMinima:	2	2	2	2	2	2	2	2	2	2	2	2	2
cotaMaxima:	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
cantMinPalabrasEnTema:	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	5.0	5.0	5.0	5.0	5.0
cotaInfAceptacionSobrante:	65.0	65.0	65.0	65.0	80.0	80.0	80.0	80.0	80.0	80.0	80.0	1.0	1.0
cotaSupDescarteTema:	20.0	13.0	20.0	11.0	11.0	11.0	11.0	11.0	11.0	11.0	11.0	80.0	80.0
cotaMinima:	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	11.0	11.0
Filtra palabras de usuario	0	0	0	0	0	0	0	0	1	1	1	1	1
cotaInfAceptacionTema:	65.0	40.0	65.0	35.0	35.0	35.0	35.0	35.0	20.0	20.0	20.0	20.0	20.0
Distancia Máxima	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0

Tabla 8.1. Parámetros de las corridas masivas

En la tabla 8.1 se describen los parámetros usados en cada corrida masiva.
maxPalsBusqueda es la cantidad de palabras de búsqueda seleccionadas de las palabras destacadas de la conversación, según la definición 4.6.
frecuenciaMinima define la frecuencia a partir de la cual las palabras son aceptadas e insertadas en el grafo de representación según lo explicado en la sección 4.1.4.
cantMinPalabrasEnTema es la cantidad mínima de palabras aceptadas en un tema, según la definición 5.1.
cotaSupDescarteSobrante es la cota superior de descarte de Sobrantes según lo explicado en la sección 5.2.2.1.
cotaInfAceptacionSobrante es la cota inferior de aceptación de sobrantes según lo explicado en la sección 5.2.2.1.
cotasupdescartetema es la cota superior de descarte de tema según lo explicado en la sección 5.2.2.1.
cotaInfAceptacionTema es la cota inferior de aceptación de tema según lo explicado en la sección 5.2.2.1.
cotaMinima es la cota inferior de normalización de los pesos, según lo explicado en la observación 4.1.
cotaMaxima es la cota superior de normalización de los pesos, según lo explicado en la observación 4.1.
Filtra palabras de usuario puede tener dos valores: 1 = filtra, 0 = no filtra. Cuando vale 1, las apariciones de nombres de usuarios dentro de los párrafos son filtradas.
Distancia Máxima es la máxima distancia representada dentro del grafo según lo explicado en la sección 4.1.4.

Observación 8.1
Los casos para los cuales se encontraron temas relacionados, estos fueron siempre subtemas del tema buscado. Esto es debido a la naturaleza de las conversaciones analizadas, y no a la lógica implementada, no siendo este hecho relevante para las pruebas de la aplicación.

8.2.1 Corrida masiva 1

Se ejecutó la aplicación con parámetros establecidos arbitrariamente.

Hay muchos temas que no son reconocidos. A pesar de que se encontró la cantidad mínima de palabras en tema, no se cumple con la cota superior de descarte de temas siendo necesario realizar ajustes sobre los parámetros a la espera de obtener un mejor comportamiento de la herramienta. La siguiente estrategia que se aplica es disminuir la cota superior de descarte de temas.

8.2.2 Corrida masiva 2

Se disminuye la cota superior de descarte de temas y la cota inferior de aceptación de temas esperando que con ésta disminución de la exigencia impuesta por éstas cotas haya más conversaciones con temas relacionados.
La cantidad de temas desconocidos fue menor, y aumentó la cantidad de temas relacionados. Se demostró un comportamiento de acuerdo a lo esperado.

8.2.3 Corrida masiva 3

Se disminuye la cantidad de *palabras de búsqueda* (definición 6.3.) a 16 (antes era 25) para ver si con menos palabras se puede mantener la bondad en los resultados.
La cantidad de temas desconocidos y temas no encontrados fue mayor, y disminuyó la cantidad de temas relacionados. La estrategia no fue buena.

Conclusión
A partir de estos resultados se deduce que es necesario una mayor variedad de palabras elegidas para la búsqueda y comparación del tema para el universo de temas en consideración.

8.2.4 Corrida masiva 4

Se incrementó la cantidad de *palabras de búsqueda* a 35 (anteriormente era 16). Esto se acompañó con una nueva disminución en los valores de la cota superior de descarte de temas y la cota inferior de aceptación de temas, disminuyendo nuevamente la exigencia sobre éstos parámetros.
La cantidad de temas desconocidos disminuyó considerablemente. Se demostró un comportamiento de acuerdo a lo esperado.

8.2.5 Corrida masiva 5

Se aumentó la cota superior de descarte de sobrantes y levemente la cota inferior de aceptación de sobrantes.
Se presentaban casos de posibles fraudes (tema aceptado y sobrantes excedidos) que no eran tales. Esto es debido a que es necesario ajustar nuevamente la cota inferior de aceptación de sobrantes.

8.2.6 Corrida masiva 6

Se filtraron las palabras de mayor frecuencia entre las distintas conversaciones: "PAIS", "ARGENTINA", "CAMBIAR", "TIEMPO", esperando que en lugar de las mismas se destaquen otras palabras destacadas de cada tema y mejoren los resultados.
Los resultados no variaron sustancialmente.

8.2.7 Corrida masiva 7

Se incorporaron las palabras destacadas de cada grupo de temas como pertenecientes a un tema que represente el grupo, y se analizan todas las conversaciones. Para esto se eligieron para cada tema, las primeras 20 palabras más frecuentes dentro de cada grupo de temas.

Conclusión

- Como se obtuvo un índice de error del 0%, estos resultados avalan que la herramienta pueda deducir, a partir de las relaciones con temas existentes y datos estadísticos, cual es el conjunto correcto de palabras que deben determinarse como relacionadas a cada tema genérico en la base de conocimiento.
- Si interpretamos el resultado de este proceso como el cálculo de una media entre las distintas conversaciones de un mismo tema, los resultados expresan que el corrimiento de cada conversación respecto de esta media es ínfimo para los parámetros establecidos.
- Además, el hecho de que las conversaciones analizadas caigan dentro de esta media expresa que el método estratégico utilizado para la distinción del tema de la conversación (patrones de comportamiento de palabras y frases), y que es el que determina esta media, es apropiado. De otra forma, las conversaciones podrían diferir mucho de esta media calculada.

8.2.8 Corrida masiva 8

Al encontrar un tema con incertidumbre, y pedir el nombre de otro tema, se asocian los sobrantes y coincidentes, no sólo los coincidentes.

La cantidad de temas encontrados con incertidumbre aumentó de manera muy significativa, y disminuyó la cantidad de temas relacionados. De esta manera, se acortan las distancias entre las conversaciones y los temas de comparación.

8.2.9 Corrida masiva 9

Se aumenta la exigencia de palabras minimas en tema a 5 (antes era 4). Se disminuye la exigencia de coincidentes para lograr que se asocie el tema con certeza (cota inferior de aceptación de tema) a 20, antes estaba en 35. Se aumenta la cota cota superior de descarte de sobrantes a 50.

Cambió el tipo de relación definida entre la conversación y el tema. Un gran número de conversaciones que antes eran asociadas con incertidumbre, ahora son relacionadas como subtemas. Es decir, las modificaciones cambian la interpretación que se les da a las relaciones entre los temas según las situaciones presentadas.

8.2.10 Corrida masiva 10

Se aumenta la cota superior de descarte de sobrantes a 65, esperando de esta manera que haya un incremento en la cantidad de temas encontrados con certeza.
La cantidad de temas encontrados con certeza aumentó significativamente.

Conclusión
Los intervalos de aceptación y rechazo deben amoldarse a la situación dada por el universo de temas utilizado. En particular, el intervalo de rechazo de sobrantes también depende de la cantidad de palabras de búsqueda sumado a hasta cuántos sobrantes son tolerados al encontrarse un tema.

8.2.11 Corrida masiva 11

En forma análoga a la corrida masiva 7, se incorporan las palabras destacadas de cada grupo de temas como pertenecientes a un tema que represente el grupo, y se analizan todas las conversaciones. Para esto se eligirá para cada tema, las primeras 35 palabras más frecuentes dentro de cada grupo de temas.

El resultado es óptimo. La cantidad de temas encontrados con certeza es del 95%. El 5% restante corresponde a temas relacionados.

8.2.12 Corrida masiva 12

Se elige la opción de filtrar los nombres de los usuarios pertenecientes a la conversación, con las mismas precondiciones de la corrida masiva 10.
No se sufren alteraciones en el comportamiento (respecto de la corrida masiva 10).

Conclusión
Se puede observar que la mención de los usuarios no inciden en la decisión tomada por la aplicación, para las conversaciones analizadas. Sin embargo, debe reconocerse que puede haber situaciones en las que los nombres de usuarios, debido a la forma de expresarse, sean mencionados reiteradamente, afectando los resultados. Filtrando los nombres nos aseguramos que esto no suceda.

8.2.13 Corrida masiva 13

Se procede en forma análoga a la corrida 11, pero esta vez dejando afuera una conversación de cada tema. Luego se carga cada tema con las 35 palabras más destacadas de su grupo. A continuación, se corren las 10 conversaciones restantes (las que quedaron afuera). Se intenta observar la manera en que responde la herramienta alimentada con información estadística, frente a conversaciones no analizadas previamente de temas conocidos.

Las variables muestran el comportamiento del análisis de las 10 conversaciones excluídas del proceso de aprendizaje estadístico. Los resultados son aceptables. El 70% fueron temas relacionados, el 20% fueron temas encontrados con incertidumbre, y el 10% restante fueron temas encontrados mal con incertidumbre. Adaptando los parámetros se puede conseguir que los temas relacionados pasen a ser temas encontrados.

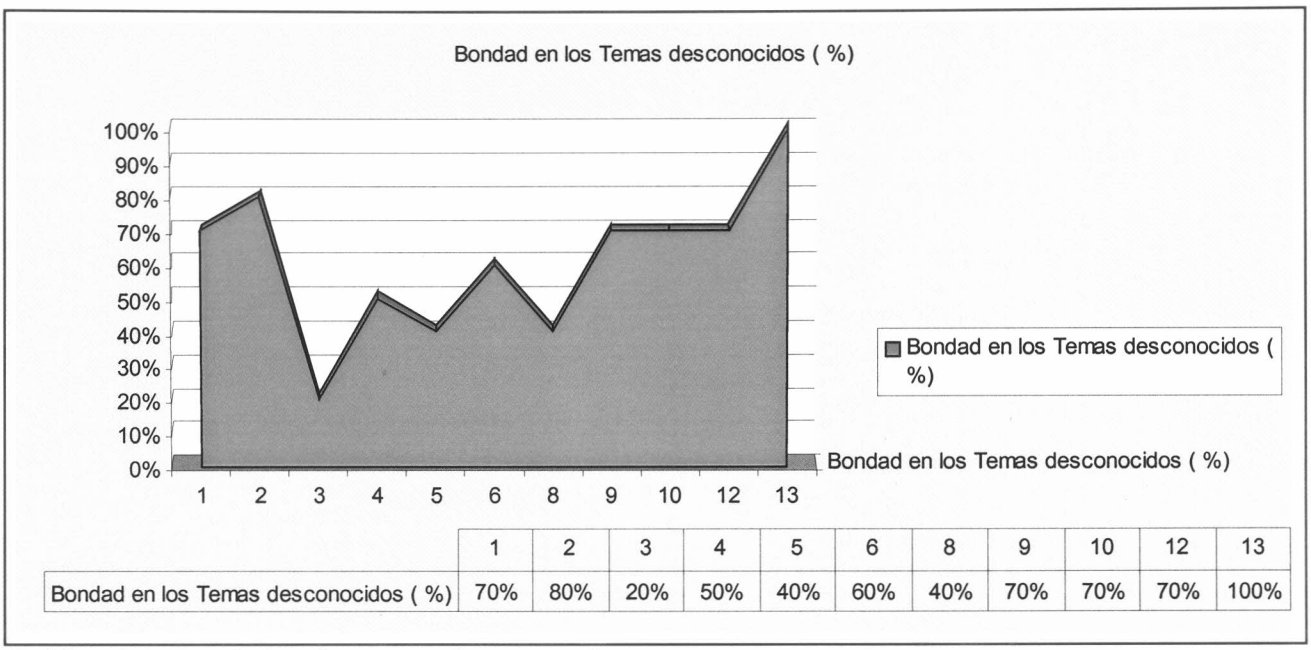
8.3 Comportamiento de las variables analizadas

Estadística / Corrida masiva	1	2	3	4	5	6	7	8	9	10	11	12	13
Temas desconocidos	13	8	18	5	4	6	0	4	7	7	0	7	0
Temas no encontrados	4	1	7	0	0	0	0	0	1	1	0	1	0
Temas encontrados	0	0	0	0	0	0	0	0	1	6	38	6	0
Temas encontrados con incertidumbre	21	21	17	26	19	20	5	19	4	4	0	4	2
Temas encontrados mal con incertidumbre	6	8	4	8	9	8	0	9	4	4	0	4	1
Temas encontrados mal con certeza	0	0	0	0	0	0	0	0	0	0	0	0	0
Temas relacionados	0	3	1	1	8	6	35	8	21	16	2	16	7
Temas mal relacionados	0	0	0	0	0	0	0	0	3	3	0	3	0

Figura 8.1. Cuadro con detalle de variables de resultados en las distintas corridas masivas

Observación 8.2: Los casos de temas mal encontrados con incertidumbre son sólo malos en el sentido de que no caen dentro del mismo conjunto de tema, pero en realidad la herramienta lo relaciona certeramente, ya que los casos son de conversaciones de *pobreza* en los cuales se habla de *inseguridad* y *economía*. Es aceptable entonces esperar que a algunas de dichas conversaciones de *pobreza* se las distinga por alguno de los temas de *inseguridad* o *economía*.

8.3.1 Temas Desconocidos

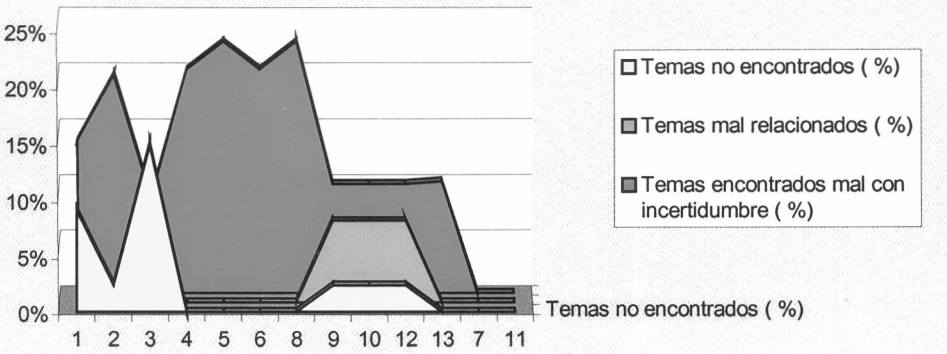


	1	2	3	4	5	6	8	9	10	12	13
Bondad en los Temas desconocidos (%)	70%	80%	20%	50%	40%	60%	40%	70%	70%	70%	100%

Gráfico 8.1. Bondad en los temas Desconocidos.

La cantidad de temas desconocidos deseables debe ser equivalente a la cantidad de temas distintos que existan en la muestra. Un exceso de temas desconocidos muestra que conversaciones a las cuales se les debería haber encontrado el tema y por el contrario esto no ocurrió.
Un valor menor que la cantidad de temas distintos existentes demuestra que hubo conversaciones que fueron relacionadas de alguna forma a temas no correspondientes.
El gráfico muestra los valores de bondad calculando el porcentaje de decisiones correctas respecto del total de temas.
Los mejores valores se encuentran en la prueba masiva 13 (con carga previa de información estadística), le sigue la prueba masiva 2, con un conjunto particular de parámetros, y luego las pruebas masivas 9,10 y 12.
El peor resultado se muestra en la prueba masiva 3, donde la cantidad de palabras de búsqueda fue disminuída bruscamente.

Errores en la decisión

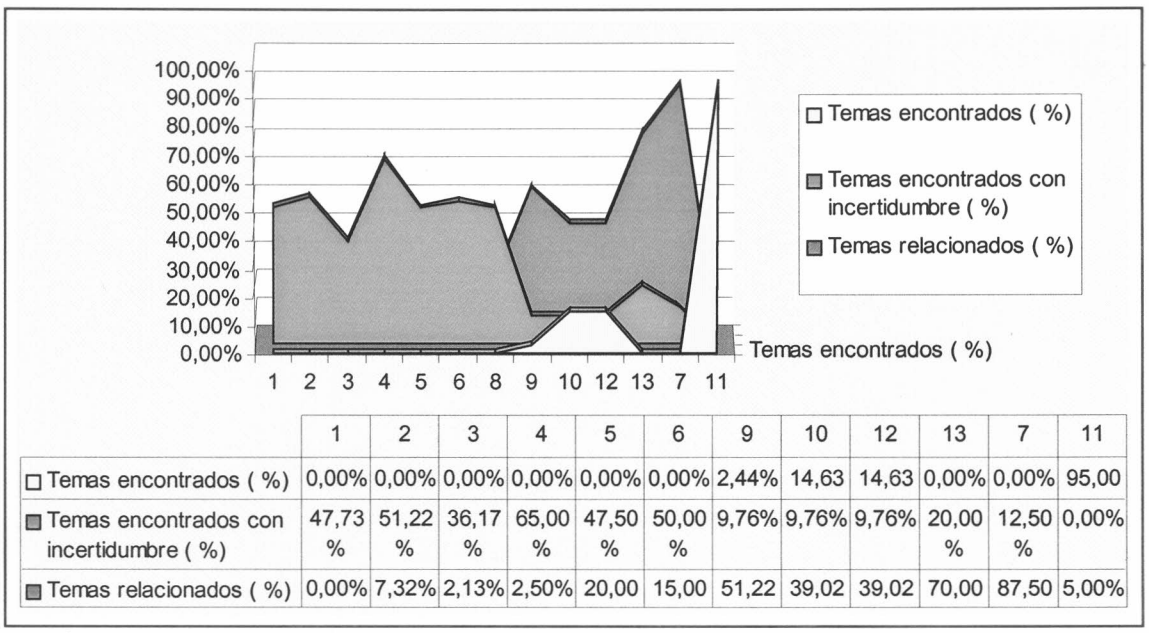


	1	2	3	4	5	6	8	9	10	12	13	7	11
Temas no encontrados (%)	9%	2%	15%	0%	0%	0%	0%	2%	2%	2%	0%	0%	0%
Temas mal relacionados (%)	0%	0%	0%	0%	0%	0%	0%	7%	7%	7%	0%	0%	0%
Temas encontrados mal con incertidumbre (%)	14%	20%	9%	20%	23%	20%	23%	10%	10%	10%	10%	0%	0%

Gráfico 8.2. Errores en la decisión.

En el gráfico 8.2. se muestran las variables que expresan distintos errores en la decisión. En la corrida 13 se puede ver que solo se obtiene un 10% de error en la decisión, viendo una mejora significativa dada por el aporte de información estadística, contra las demás corridas, donde los valores totales oscilan entre el 17 y el 24 por ciento de las decisiones. Dependiendo del grado de gravedad que se le asigne a cada uno de los distintos tipos de mala decisión se puede deducir, de las demás corridas (quitando la corrida masiva 13) cual expresa la situación mas conveniente.
Las corridas 7 y 11, basadas en información estadística del total de las conversaciones previamente cargada, no muestran ningún tipo de error en las decisiones tomadas.

8.3.2 Aciertos en la decisión

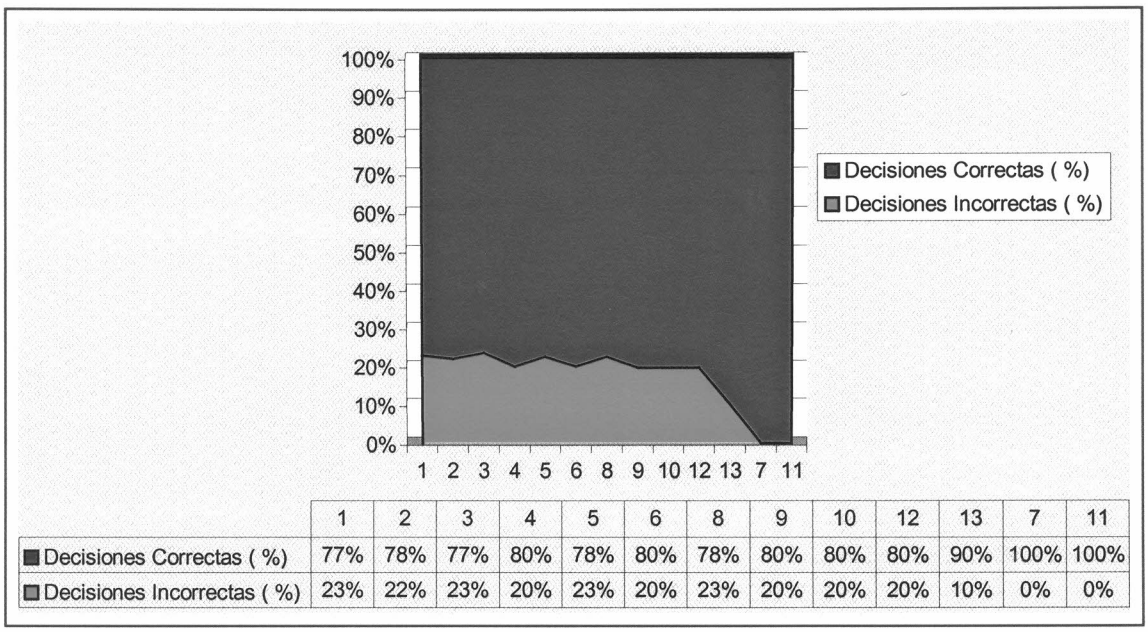


	1	2	3	4	5	6	9	10	12	13	7	11
Temas encontrados (%)	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	2,44%	14,63	14,63	0,00%	0,00%	95,00
Temas encontrados con incertidumbre (%)	47,73 %	51,22 %	36,17 %	65,00 %	47,50 %	50,00 %	9,76%	9,76%	9,76%	20,00 %	12,50 %	0,00%
Temas relacionados (%)	0,00%	7,32%	2,13%	2,50%	20,00	15,00	51,22	39,02	39,02	70,00	87,50	5,00%

Gráfico 8.3. Aciertos en la decisión.

En este gráfico se pueden ver los distintos tipos de decisiones que fueron tomadas correctamente. Se le da una mayor importancia a las decisiones que condujeron a la determinación en forma certera del tema de la conversación. Las corridas 10 y 12 son las que mejor incidencia de este tipo tienen. La corrida 7 y 11 (basadas en carga estadística previa del total de las conversaciones) muestran una mayoritaria proporción de temas relacionados y temas encontrados respectivamente. La corrida 13 (carga estadística previa), a pesar de no tener temas encontrados en forma certera, es la que mayor porcentaje de decisiones correctas tiene (90 por ciento).

Proporción de Decisiones Correctas y Decisiones Incorrectas



	1	2	3	4	5	6	8	9	10	12	13	7	11
Decisiones Correctas (%)	77%	78%	77%	80%	78%	80%	78%	80%	80%	80%	90%	100%	100%
Decisiones Incorrectas (%)	23%	22%	23%	20%	23%	20%	23%	20%	20%	20%	10%	0%	0%

Gráfico 8.4. Proporciones de Decisiones Correctas y Decisiones Incorrectas.

En el gráfico 8.4. se compara para cada corrida las proporciones entre las decisiones tomadas correctamente (temas encontrados, temas encontrados con incertidumbre y temas relacionados), marcados con azul, y las decisiones tomadas en forma incorrecta (temas no encontrados, temas mal relacionados y temas encontrados mal con incertidumbre), marcados con rojo.
Se puede apreciar que las mejores proporciones se obtienen en las corridas 7 y 11 (corridas especiales con carga de información estadística de todas las conversaciones). Excluyendo estas dos corridas masivas, los mejores resultados se obtienen en la corrida 13 (con carga previa de información estadística) y luego la corrida 6 (que tiene los mismos parámetros que la corrida 5, filtrando las palabras más comunes entre las distintas conversaciones).

8.4 Resumen

En este capítulo se expuso un detalle de las pruebas realizadas sobre la herramienta y los resultados obtenidos.
Se puede apreciar que los mejores resultados se obtienen cuando la base de conocimiento contiene información estadística deducida a partir de la incorporación de conocimiento de conversaciones previamente analizadas.

Según el uso que se le de a la herramienta, pueden adoptarse dos estrategias en la decisión.
La primera estrategia es determinar, siempre que sea posible, el tema de la conversación, asumiendo que se tiene un universo completo de temas en la base de conocimiento. Para esto, se eligirá como más cercano a la conversación el tema que mayor peso de palabras coincidentes tenga, sin tener en cuenta la cantidad de sobrantes. Esto es viable en la medida en que los temas en la base de conocimiento sean genéricos, y estén representados por un gran conjunto de palabras. Entonces se eligirá el tema que mayor peso de coincidentes tenga con la conversación, sin interesar la proporción de estos coincidentes sobre el total de palabras representantes de la conversación.
La otra estrategia es priorizar la proporción de coincidentes sobre el total de palabras del tema. Esta estrategia sirve para incorporar nueva información en la BC y para incorporar nuevos temas y nuevas relaciones con temas de la BC.
Para variar de una estrategia a otra, es necesario cambiar la consulta para que priorice el tema de mayor peso de coincidentes por sobre el tema de mayor puntaje (que es la proporción de coincidentes sobre el total de palabras del tema).
Si no se desea analizar las palabras sobrantes, entonces es necesario subir la cota de rechazo de sobrantes hasta el límite necesario, de modo que los sobrantes sean siempre descartados y no se pregunte un tema asociado a los mismos.

El entorno en el cual se corrieron las pruebas es una máquina tipo Pentium III de 700 Mhtz de velocidad, 128 Mbytes de memoria RAM y el servidor de la base de datos corriendo en la misma máquina que la aplicación. Se puede considerar que la aplicación tiene una buena performance en el uso del tiempo, procesando conversaciones de 10 a 14 Kilobytes de texto en no más de 20 segundos.

Capítulo 9

Conclusiones

El objetivo de la presente tesis fue la construcción de una herramienta para encontrar en forma automatizada los temas de conversaciones informales, generalmente conocidos como "chats". Fue necesario descubrir y estudiar características comunes en la forma de expresión de los chats, para poder determinar a partir de las mismas el tema de la conversación. También se necesitó definir formas de representación adecuadas para la conversación en estudio y para la información almacenada en la base de conocimiento que es utilizada por la herramienta para recuperar información para la deducción del tema y para almacenar nuevo conocimiento incorporado por el usuario o deducido en forma automática.

Se exponen a continuación conclusiones relacionadas al trabajo en general dignas de mención y se proponen algunos temas de investigación y desarrollo relacionados.

9.1 Características de la solución propuesta

- Los valores de las Cotas usadas por la aplicación dependerán del uso particular que se dé a la herramienta, ya que, como se analizó en forma teórica, las cotas son dependientes del universo de temas en consideración. Y dentro de este universo, dependerán de las relaciones entre los distintos temas.

- La aplicación soporta distintos universos temáticos. El universo puede estar compuesto sólo por temas generales, solamente temas específicos o ambos. Para cada situación habrá una composición distinta de la base de conocimiento que se adapte a la necesidad.

- La incorporación de errores comunes de tipeo y ortografía en el Diccionario Normalizador es una solución sencilla para un problema importante que es la mala escritura presente en las conversaciones informales, ayudando de esta manera a reconocer palabras que puedan ayudar a distinguir el tema de la conversación.

- El conocimiento estático puede servir para estudiar alguna situación en particular, o estudiar características de un grupo de conversaciones analizadas, pero el mismo es

insuficiente para ser usado por un Sistema Experto para el reconocimiento del tema de conversación.

- En el análisis de charlas o conversaciones informales se presenta el análisis sintáctico como un inconveniente, debido a que las conversaciones de este tipo no respetan estructuras sintácticas. Es así que se orientó la búsqueda hacia el estudio de los patrones de comportamiento de las palabras y frases. Sin embargo, este análisis no es restrictivo al dominio de conversaciones informales, sino que también puede aplicarse para el análisis de distintos tipos de conversaciones formales y narraciones, es decir el análisis de textos generales, sin importar su estructura.

- Si los temas existentes en la base de Conocimiento son bien distintos entre sí, entonces la herramienta se puede parametrizar para que, con sólo encontrar una pequeña cantidad de palabras de algún tema, el mismo sea elegido como el tema relacionado a la conversación. A su vez, cada tema en la base de conocimiento tendrá asociado un amplio conjunto de palabras. Si los temas son parecidos entre sí, es decir que tienen una considerable cantidad de palabras en común, entonces debe haber una mayor exigencia de palabras coincidentes entre la conversación y el tema a comparar, para así distinguir entre distintos temas parecidos.

- La aplicación es capaz de identificar temas y relaciones con temas conocidos solamente a partir de la información asimilada en el autoaprendizaje.

- Inferencia Estadística
  Se ha visto que una palabra o frase será calificada como clave si la probabilidad de que se hable de un tema en especial cada vez que aparece esta palabra es muy alta, además de otras condiciones. Esto puede suceder con un tema o con un grupo de temas. Calcular la probabilidad de que dicha palabra se refiera a uno u otro tema en teoría es bastante sencillo si se piensa que esto puede hacerse a partir de una muestra considerable de conversaciones de los temas involucrados. Si esto sucede para una palabra sobre muchos temas, entonces puede deducirse algún tipo de probabilidad asociada a cada uno de estos temas para la palabra en cuestión. Esto se logra contando la cantidad de temas para los cuales está dicha palabra y en los cuales se califica a la misma como clave, y calculando la proporción de la misma en cada tema. Esta probabilidad puede ir corrigiéndose a través del tiempo, con el análisis e incorporación de nuevas conversaciones asociadas al mismo tema o grupo de temas de las cuales esa palabra era considerada clave.
  De la misma forma, se puede hacer un análisis de con qué porcentaje de palabras coincidentes el usuario acepta los temas ofrecidos por la aplicación, para de esta forma efectuar pequeños corrimientos en los parámetros asociados a las cotas de aceptación de temas. Así también pueden inferirse a través de la interacción con el usuario valores más certeros para cada una de las cotas asociadas con los intervalos de aceptación, incertidumbre y rechazo de los temas. Una palabra con una alta concentración puede ser interpretada como una palabra que sirve de estímulo a la conversación. Posiblemente si esta palabra ayuda al reconocimiento del tema, pueda ser candidata a ser una clave para dicho tema. Hay que tener en cuenta que esta puede ser una forma para identificar

palabras claves de temas que no se tienen conocimiento almacenado. Ejemplo: un usuario identifica a la conversación con un tema todavía no conocido, entonces podrá inferirse que son claves las palabras que sigan este patrón (alta concentración).

- Si la base de conocimiento se alimentara en forma constante con la información surgente del análisis estadístico, los temas podrían ser reconocidos a partir de esta información acumulada previamente. Entonces, ¿ Qué ventaja tendría en este contexto el análisis de patrones?. En principio, el análisis de comportamiento sumado al autoaprendizaje sirven para obtener buenos resultados con poca o nula información estadística. En las pruebas realizadas se observa que la curva de aprendizaje crece pronunciadamente al iniciar el uso de la herramienta (se deducen temas o temas relacionados con muy pocas conversaciones analizadas previamente), y se espera que se nivele a medida que se acreciente la cantidad de información almacenada en la base de conocimiento. Además, la información estadística se basará en estos patrones de comportamiento, y no meramente en la presencia o no de determinadas palabras en una conversación.

9.2 Temas abiertos

- El parser puede adaptarse para trabajar en la detección de frases comúnmente mencionadas, de tal forma que las mismas puedan ser identificadas en forma automática, aportando así información adicional a las conversaciones.

- Los chats tienen un problema, y es que muchas veces hay usuarios infiltrados que intervienen para hablar de temas distintos a los del foro original, o se producen excepciones donde el tema principal se desvirtúa y por algún "momento" se pasa a hablar de otro tema, como ser, por ejemplo, algo vinculado a las noticias del día. En estas situaciones la herramienta puede adaptarse para cumplir con las tareas administrativas de un moderador, reagrupando las frases de otros temas en los foros que correspondan, y dando aviso de estos cambios a los foristas involucrados.

- Puede adaptarse la herramienta para que, sin necesidad de conocimiento previo, organice y clasifique documentos según las características deducidas a partir de las palabras y frases y el comportamiento de las mismas dentro de cada documento en estudio. De la misma forma puede adaptarse para la clasificación de mensajes de correo electrónico.

- El descubrimiento y análisis de nuevos patrones es fundamental para el enriquecimiento de la aplicación con el propósito de lograr una mayor precisión en los resultados.

- Puede desarrollarse un análisis morfológico de los grafos generados para cada conversación, y ver como puede usarse este conocimiento adicional para la determinación del tema, tal vez para calcular la proporción de cada uno de los distintos temas que se hablan en la conversación, y su distribución a partir de las distintos subgrafos identificables, o la formación de subgrafos disconexos, etc.

Bibliografía

[BN00] Advances in Probabilistic and Other Parsing Technologies (Text, Speech and Language Technology, Volume 16) by Harry C. Bunt, Anton Nijholt, October 2000, Kluwer Academic Publishers; ISBN: 0792366166.

[ND97] Computer Based Learning Unit, Nikos Drakos, University of Leeds, 1997.

[WB90] Designing Object-Oriented Software, Rebeca Wirfs-Brock, 1990. Prentice Hall; ISBN: 0136298257.

[EG97] Design Patterns, Erick Gamma, 1997. Addison-Wesley Pub Co; ISBN: 0201633612.

[DFLDH88] Using latent semantic analysis to improve access to textual information; S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester and R. Harshman; *Conference proceedings on Human factors in computing systems*, 1988, Pages 281 – 285.

[RK94] Inteligencia Artificial, Elaine Rich y Kevin Knight, 1994, McGraw-Hill; ISBN: 0-07-052263-4.

[IFS00] Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems,by Guanrong Chen, Trung Tat Pham, November 2000).Lewis Publishers, Inc.; ISBN: 0849316588.

[DK79] Logic and semantic networks, Amaryllis Deliyanni y Robert A. Kowalski, ACM 1979. Digital library.

[JA87] Natural language understanding, James Allen,1987. Addison-Wesley Pub Co; ISBN: 0805303340.

[RG82] Three principles of representation for semantic networks, Robert L. Griffith, ACM 1982.

[DL90] Parsing semantic dependencies in associative networks, Dekang Lin, ACM 1990. Digital library.

[SCH82] Relating sentences and semantic network with procedural logic, Robert F. Simmons y Daniel Chester, ACM 1982.

[SA00] Conversation Map: A Content-Based Usenet Newsgroup Browser. Warren Sack – MIT Media Laboratory, Julio 2000.

[SA01] Conversation MAP, an interface for very large-scale conversations. Warren Sack – *MIT Media Laboratory, Enero - 2000.*

[KSZE88] A freely Available Wide Coverage Morphological Analyser for English, Daniel Karp, Yves Schabes, Martin Zaidel and Dania Egedi, *Proceedings of COLING-92*, 1992.

[WD00] Incorporating a semantic analysis into a document retrieval strategy; Edgar B. Wendlandt and James R. Driscoll; *Proceedings of the fourteenth annual international ACM/SIGIR conference on Research and development in information retrieval*, 1991, Pages 270 – 279.

Indice de Ilustraciones