1 Introducción

1.1 Antecedentes

Desde los primeros estudios del lenguaje, pero más en los últimos años, la resolución de la anáfora ha sido foco de investigación de filósofos, lingüistas, científicos del conocimiento e IA (**I**nteligencia **A**rtificial), de sicolingüistas y de lingüistas computacionales [Mitkov, 98a]. Su importancia radica, entre otras razones, en que la anáfora:

- es uno de los fenómenos más complejos dentro del lenguaje natural [Huang, 2000; Mitkov, 2001]
- es considerada uno de los problemas fundamentales de la lingüística y Chomsky se apoya en ella, para mantener la teoría de que la facultad del habla es innata [Chomsky, 1986]
- se ha demostrado que en ella interactúan factores sintácticos, semánticos y pragmáticos [Hirst, 1981; Huang, 2000]
- es necesaria en un amplio rango de tareas del PLN (**P**rocesamiento del **L**enguaje **N**atural) como interfaces en lenguaje natural, la comprensión del lenguaje, traducción automática, extracción de información y generación automática de resúmenes [Hirst 1981; Fox 1987; Cornish 1996; Fretheim y Gundel 1996; Hahn et al 1996; Kameyama 1997; Mitkov, 2001]

La anáfora indirecta establece un enlace asociativo entre una entidad lingüística (palabra o expresión) con alguna entidad implícita introducida previamente a través del texto en el discurso. En las últimas décadas ha recibido una especial atención dentro de diferentes disciplinas que la han tratado desde varias perspectivas. Así, dentro de la tradición lingüística, se encuentran Erkü y Gundel [1987], Huang [1994] y Matsui [1995]; dentro de la sicolingüística, están Clark y Haviland [1977], y Sanford y Garrod [1981]; dentro de la Inteligencia Artificial, está Sidner

[1983]; y finalmente dentro de la lingüística computacional están Murata y Nagao [1996] y Gelbukh y Sidorov [1999].

La ***anáfora indirecta,*** mencionada por primera vez en el trabajo de Chafe [1976], ***es uno de los casos más difíciles de relación anafórica*** [Erkü y Gundel, 1987; Kempson, 1982; Matsui, 1995; Huang, 1994; Murata y Nagao, 2000]. También conocida como: conexión referencial [Clark, 1977], anáfora asociativa [Hawkins, 1978], anáfora inferenciable [Prince, 1981], anáfora implícita u oculta [Sidorov y Gelbukh, 1999], conexión de referencia cruzada [Huang, 2000], ***ha sido un caso poco abordado*** en la lingüística computacional **a pesar de** ***su importancia para determinar la coherencia del texto*** [Mitkov, 2001]. Aumentar el conocimiento sobre ella, las condiciones que la determinan, sus mecanismos de procesamiento y dotar de ellos a la computadora para apoyar el PLN son las metas inmediatas a lograr en este trabajo.

1.2 Situación actual

Se podría sintetizar rápidamente con el comentario de que la mayoría de los problemas de la lingüística computacional se presentan en la resolución de la anáfora indirecta, debido a la necesidad de hacer explícita a la computadora toda la información y relaciones requeridas para "*entender*" el texto. Con el fin de explicar lo anterior, se hará un rápido recorrido de las necesidades y los problemas encontrados en el PLN que de una u otra forma afectan la resolución de la anáfora y al final se abordará la anáfora indirecta en particular.

1.2.1 La anáfora

La mayoría del trabajo anterior en la resolución de la anáfora ha utilizado mucho conocimiento del dominio y lingüístico, además de requerir considerable captura manual, [Sidner, 1979; Carbonell y Brown, 1988; Rich y Luperfoy, 1988] lo que ha dificultado la representación y procesamiento. Sin embargo, la necesidad apremiante para desarrollar sistemas robustos y menos costosos ha llevado a los investigadores, a partir de 1990, a alejarse un poco del conocimiento lingüístico e intentar estrategias de solución que requieran menor conocimiento (Knowledge-poor) [Dagan e Itai, 1990; Kennedy y Bougarev, 1996; Baldwin, 1997; Mitkov,

1998a, 2000b y 2000c]. Se han utilizado además estrategias combinadas para la resolución de la anáfora en el Español [Palomar et al, 2001].

La posibilidad de corpus sin etiquetar y de corpus etiquetados con enlaces referenciales, dio un fuerte impulso a la resolución de la anáfora tomando en cuenta el entrenamiento y la evaluación; los corpus (especialmente cuando están etiquetados) son un recurso de gran valor para la investigación empírica y los métodos de aprendizaje automático que animan el desarrollo de diferentes enfoques, posibilitando también medios para la evaluación de algoritmos desarrollados. Desde simples reglas de co-ocurrencia [Dagan e Itai, 1991] pasando por el entrenamiento de árboles de decisión para identificar las parejas anáfora y antecedente [Aone y Bennett, 1995], hasta algoritmos genéticos para optimizar los factores que afectan la resolución de la anáfora [Orasan et al, 2000], han sido logrados gracias a la posibilidad de contar con corpus adecuados.

El preprocesamiento del texto, o procesamiento previo a la aplicación del método de resolución de la anáfora a un texto, es un problema significativo ya que la exactitud es demasiada baja y como consecuencia el rendimiento de estos sistemas está lejos del ideal; la dependencia vital del sistema de resolución de la anáfora es tal que tendrá poco rendimiento, aunque el método sea muy bueno. En esta etapa, *los problemas principales* se encuentran en el análisis morfológico, etiquetado de partes de la oración, reconocimiento de entidades nominales, reconocimiento de pronombres, reconocimiento de palabras desconocidas, extracción de frases nominales, descomposición analítica (parsing), etc. Por ejemplo: la mejor exactitud encontrada para la descomposición analítica de textos sin restricción es alrededor del 87% [Collins, 1997]; el mejor rendimiento logrado con etiquetadores de entidades nominales da una exactitud del 96% cuando se prueba y utiliza en corpus con *noticias* sobre un dominio o tema específico [Mitkov, 2001].

Como resultado de las limitantes mencionadas, *la mayoría de los sistemas de resolución de la anáfora no operan de modo totalmente automático, y algunos métodos han sido simulados sólo manualmente*. Como ejemplos ilustrativos: la resolución propuesta por Hobbs no fue implantada en su versión original [Hobbs, 1976, 1978]; hay trabajos donde se corrigieron manualmente los resultados de las etapas de preprocesamiento (para poder utilizarlas en el algoritmo de resolución anafórica) [Lappin, 1994; Ferrandez et al, 1997; Mitkov, 1998b];

finalmente hay trabajos donde se utilizaron corpus etiquetados manualmente sin etapa de preprocesamiento [Ge et al, 1998; Tetreault, 1999]. Reaccionando a la situación mostrada se han iniciado esfuerzos, a largo plazo, para lograr sistemas totalmente automatizados [Fukumoto et al, 2000; Tanev y Mitkov, 2000; Mitkov, 2001].

1.2.2 La anáfora indirecta

De los modelos de análisis de la anáfora indirecta, tres son los que más influencia han tenido en el área:

- El modelo de relevancia
- El modelo de tópico o focal
- El modelo de escenario

1.2.2.1 El modelo de relevancia

La teoría de relevancia supone que el mecanismo cognitivo central del ser humano es un dispositivo deductivo generador de inferencias que trabaja tratando de maximizar la relevancia con respecto a la comunicación; el principio de relevancia es el responsable de recuperar el contenido explícito o implícito de un enunciado. En otras palabras, al interpretar un enunciado **el receptor** estará siempre "*maximizando los efectos contextuales del enunciado*" y "*minimizando los esfuerzos de procesamiento del enunciado*" [Matsui, 1995]. Dentro de este modelo, Kempson [1988a] observa que la interpretación de la anáfora indirecta requiere un análisis semántico / pragmático, más que gramatical, y de la información asociada con premisas adicionales implícitas.

En el modelo de relevancia, la idea básica es que la interpretación de la anáfora indirecta se encuentra *suponiendo conexiones* que se apoyan en efectos contextuales apropiados pero *sin sujetar el lenguaje a esfuerzos injustificados* para obtener estos efectos. La validez del análisis de la anáfora indirecta, y de la anáfora en general, con el modelo de relevancia depende crucialmente de cómo se aplica el principio, o más concretamente de cómo se pueden obtener y balancear tanto los "efectos contextuales" como los "esfuerzos de procesamiento". Desgraciadamente, en los trabajos consultados [Matsui, 1993, 1995; Kempson, 1988a, 1988b;

Sperber y Wilson, 1995; Levinson, 1989] no existe un mecanismo satisfactorio para medir el balance costo-beneficio; no parece que el principio de relevancia haya podido ser implantado confiablemente; y se ha reportado dificultad empírica al probarlo [Huang, 2000].

1.2.2.2 El modelo de tópico o focal

En el modelo de tópico, la idea básica es que la interpretación de la anáfora indirecta está determinada principalmente por el tema o tópico (aquello sobre lo que se está hablando) de las oraciones previas del discurso. Este enfoque está representado por los trabajos de Sidner [1983] y Erkü y Gundel [1987]. La interpretación de la anáfora indirecta se efectúa por un algoritmo que selecciona el foco del discurso con base en un conjunto ordenado de preferencias; además, las interpretaciones resultantes del algoritmo quedan sujetas a los requerimientos de consistencia con el conocimiento del mundo.

1.2.2.3 El modelo de escenario

En el modelo de escenario, la idea básica es que la interpretación de la anáfora indirecta se encuentra siempre referida a un dominio mental apropiado de referencia. Este enfoque está representado por el trabajo de Sanford y Garrod [Sanford y Garrod, 1981; Garrod y Sanford, 1994]. Apoyándose en nociones como: marcos [Minsky, 1975; Fillmore, 1982], esquemas [Rumelhart, 1980; Chafe, 1987] y de guiones [Schank y Abelson, 1977], denominaron a este dominio de referencia ***un escenario***. Un escenario, de acuerdo a Sanford y Garrod, puede ser activado desde tres dimensiones: ***actual***, porque se encuentra en el foco de atención del receptor y almacenado en la memoria primaria, ***o antigua***, si no es parte del foco de atención del receptor y se encuentra en la memoria secundaria; ***explícito***, se refiere a las entidades que han sido mencionadas directamente en el discurso, ***o implícito***, son entidades que no han sido explícitamente mencionadas pero que están relacionadas en forma relevante con algo mencionado en el discurso; ***de entidad***, representada por los individuos que son los principales protagonistas de una escena, ***o de rol***, representada por los papeles representados en el escenario descrito en el discurso.

1.2.2.4 Comparación de modelos

Para observar las ventajas y desventajas de los tres modelos se puede intentar una comparación manual tomando como base el ejemplo de Erkü y Gundel [1987] y analizándolo desde cada enfoque.

(1) Juan entró a un **restaurante**. El *mesero* era italiano.

En el enfoque focal el *restaurante* es el foco del discurso y por lo tanto el antecedente del *mesero*. Dentro del marco de relevancia, la suposición de conexión de que el *restaurante* donde Juan entró tiene al menos un *mesero* proviene de la extensión del contexto por el conocimiento enciclopédico (del sentido común); como consecuencia toda la interpretación es consistente con el principio de relevancia. Finalmente, en el enfoque de escenario, el uso de *restaurante* invoca un escenario que contiene en forma implícita al menos un *mesero*.

Tomando un ejemplo, un poco más complicado [Huang, 2000].

(2) Juan se detuvo por un café en un ***bar capuchino*** antes de comer en un **restaurante**. El *mesero* era italiano**.**

Este ejemplo contiene más de un antecedente posible (un **restaurante** o un **bar capuchino**) para la anáfora indirecta el *mesero*, donde el antecedente preferido sería un **bar capuchino**; de acuerdo al conocimiento común capuchino = bar italiano y como el mesero era italiano se puede inferir que era el mesero del bar capuchino.

Esta interpretación sería correcta desde el modelo focal porque el algoritmo de Sidner tomaría un **bar capuchino** como el tópico o foco del discurso (por el orden de aparición en la primera oración). Dentro del marco de relevancia, asumiendo que un **bar capuchino** es más accesible (por ser de tipo italiano) que un **restaurante**, sería la conexión preferida para la interpretación correcta. Finalmente, en el análisis de escenario habría dos antecedentes posibles, uno el **bar capuchino** y otro el **restaurante**, cada uno de ellos con posibilidad de tener *mesero*. Aquí no hay mecanismo para escoger entre ambos escenarios y queda confuso como se puede derivar una interpretación correcta bajo este enfoque. Para finalizar las comparaciones, se analizan un par de ejemplos similares tomados de Huang [2000].

(3) Juan se detuvo por un café en un **bar capuchino** antes de visitar un **museo de instrumentos musicales**. El *mesero* era italiano.

(4) Juan se detuvo por un café en un **bar capuchino** antes de visitar **un museo de instrumentos musicales**. El *encargado* era italiano.

Intuitivamente, el primer ejemplo parece menos complejo que el segundo y el porqué se encuentra en el conjunto de factores que afectan la interpretación. Clark y Haviland [1977] han identificado: la *distancia* de la conexión (el número de suposiciones necesarias para la conexión), la *plausabilidad* de la conexión (el grado de veracidad de las suposiciones) y la *computabilidad* de la conexión (el grado de facilidad en el calculo de las suposiciones); otros factores pueden incluir la *accesibilidad* (facilidad de acceso) a los *antecedentes* y a las *suposiciones contextuales*; y la *coherencia general* del discurso [Huang, 2000; Matsui, 1995]. Bajo el enfoque focal o del tópico, el factor de *accesibilidad* a los antecedentes parece jugar un rol crucial para explicar porqué el primer ejemplo es menos complejo que el segundo: mientras el antecedente para *el mesero* en el primer ejemplo es el foco del discurso, el antecedente para *el encargado* en el segundo no lo es (puede existir *un encargado* o supervisor tanto en el museo como en el bar). Por otro lado, en el análisis de relevancia la complejidad mayor del segundo ejemplo se puede atribuir al injustificado esfuerzo de proceso que debe realizar el lector para interpretar la anáfora indirecta *el encargado* en el discurso (como resultado de la mayor *accesibilidad* focal para *un bar capuchino* y el antecedente que se quiere, *un museo de instrumentos musicales*).

Finalmente, en el modelo de escenario se tendría que utilizar la noción de *accesibilidad* a las suposiciones contextuales para detectar las diferencias entre ambos ejemplos: el antecedente para *el mesero* en el primero se encuentra en el contexto (escenario) más accesible para el lector que el antecedente para *el encargado* en el segundo, considerando, por supuesto, que exista un mecanismo para decidir en cual de los escenarios actualmente activados es más accesible.

De la comparación anterior, pueden apreciarse las razones por las cuales los trabajos encontrados para la resolución de la anáfora indirecta hayan optado por utilizar los modelos, de tópico o focal y escenario.

1.2.3 Problemas pendientes de resolver

Aunque en los últimos diez años ha habido un considerable avance en el campo de resolución de la anáfora, existen aún considerables problemas sin resolver o que requieren ser atendidos para apoyar su resolución, y que representan los mayores retos para el desarrollo futuro.

Para empezar, no se identifica claramente un solo conjunto de factores (léxico, sintáctico, semántico y pragmático) en la resolución de la anáfora y si este conjunto de factores una vez agrupados estaría completo. En general los factores son divididos en *restricciones y preferencias* [Carbonell y Brown, 1988] pero otros autores arguyen que deberían considerarse como *escala de preferencias* más o menos restrictiva llamándolas simplemente *factores* [Preuß et al, 1994], *síntomas* [Mitkov, 1995] o *indicadores* [Mitkov, 1998a].

Una vez definidos conviene ver: el impacto individual de cada factor y su secuencia o coordinación al actuar [Carter, 1990]; esclarecer la existencia o no de *dependencia* (o dependencia mutua) de los factores; verificar si son aplicables por igual a todas las lenguas o son específicos de cada lengua. Algunos autores apoyan la idea de que los factores tienen aplicabilidad general a todas las lenguas, pero que las lenguas difieren en la importancia relativa de los factores [Mitkov, 1997]; además se observa, que la diferencia se da por la evolución de las lenguas por lo que podemos hablar de lenguas donde predomina más la sintaxis que la pragmática y viceversa.

"Desde el punto de vista diacrónico, las lenguas parecen cambiar de ser más pragmáticas a más sintácticas; desde una perspectiva sincrónica, las diferentes lenguas están simplemente en diferentes etapas de este círculo evolutivo" [Huang, 2000].

La resolución de la anáfora indirecta requiere conocimiento implícito, previo o de "sentido común", aunado a un análisis pragmático; *lo que la gramática provee es meramente un conjunto de restricciones que el valor identificado de una expresión anafórica debe satisfacer* [Kempson, 1988a, 1988b]. De acuerdo a esto, la interpretación de los diferentes tipos de anáfora depende de los diferentes tipos de información disponibles al lector, por lo que la interpretación para establecer el valor de la anáfora referencial y la acotada por variable (bound-anaphor) se logra con la información que se ha *presentado previamente en el contexto lingüístico*; la

interpretación de la deixis anafórica se logra con la información *presente en el contexto del discurso*; y la interpretación de la anáfora indirecta a través de información de *conocimiento implícito en el contexto del discurso*, asociado con premisas adicionales.

1.3 Definición del problema

En el PLN se han desarrollado **sistemas** que dependen de la frecuencia de palabras (las más usadas) en el texto **que sólo establecen referencias explícitas a entidades mencionadas en el texto**, pero ¿Qué pasa con las referencias a las mismas entidades a través de fenómenos como la anáfora? **Limitan la eficacia y eficiencia de los mismos.** Como ejemplos de sistemas desarrollados con este enfoque se tienen sistemas para: encontrar los temas principales en documentos [Guzmán-Arenas, 1999] y sistemas de búsqueda temática de documentos [Alexandrov et al, 2000]. El poder identificar y resolver las referencias anafóricas aumentaría su efectividad; en otras palabras, ***para interpretar por completo el texto es necesario salvar la barrera de las referencias anafóricas, sin esto se mantiene la limitante actual en el PLN***.

Aunque en los últimos años ha habido un considerable avance en el campo de resolución de la anáfora, existen aún discusiones de carácter teórico y práctico que frenan, por así decirlo, el avance en la comprensión del lenguaje [Krahmer y Piwek, 2000]. Para avanzar en la comprensión del lenguaje natural, por medio de la verificación de la coherencia textual **es necesario investigar para obtener una definición más precisa** de la anáfora en general, y **de la anáfora indirecta** en particular, **descubriendo y estableciendo sus características distintivas que permitan elaborar un mejor modelo para implementarlo en la computadora**.

En este trabajo se continuó con la investigación iniciada por Gelbukh y Sidorov [1999] para determinar:

- las condiciones de validez en la formación y los rasgos distintivos (marcadores) que permiten *detectar la existencia posible o no de la anáfora indirecta en un texto*
- cómo debe interpretarse la anáfora indirecta
- cómo debe seleccionarse el antecedente apropiado ante la existencia posible de múltiples anáforas y antecedentes

1.4 Objetivo

El objetivo general ha sido aumentar el conocimiento sobre la anáfora indirecta, las condiciones que la determinan, sus mecanismos de procesamiento y dotar de ellos a la computadora para apoyar el PLN como contribuciones de este trabajo a la investigación, que se conduce en el mundo a largo plazo, en dos tareas: comprender como aprende el ser humano el lenguaje y construir programas que permitan a la computadora entenderlo; ambas están íntimamente relacionadas.

El objetivo específico de este trabajo ha sido: **desarrollar el modelo, el método, los diccionarios y el software para resolver la anáfora indirecta en los textos en español.**

1.5 Justificación

La resolución de la anáfora es fundamental para el desarrollo de interfases de lenguaje natural, principalmente en Internet que se está desarrollando aceleradamente, para poder involucrar más gente a la utilización y beneficios de la computación, de las Bibliotecas Digitales, etc., donde se requieren aplicaciones para obtener información textual de imágenes e indexarla, donde la falta de nitidez obliga a utilizar reconocimiento aproximado para apoyar el OCR [Wu et al, 1997; Morales, 1999]; se necesitan nuevos métodos para investigar y recuperar datos de la creciente información textual disponible; y es cada día más apremiante entender el texto a un nivel lingüístico más profundo para cubrir estas demandas [Uchida et al, 1999; Gelbukh, 2000].

Uno de los problemas puntuales del procesamiento de lenguaje natural es la verificación de coherencia textual y es donde la resolución de la anáfora tiene una aportación muy importante porque permitirá obtener resúmenes de forma automatizada.

1.6 Limitaciones y delimitaciones

Este trabajo identifica las condiciones en las que se da la anáfora indirecta considerando al resto de las referencias como parte de los algoritmos que se desarrollan; se ha profundizado sólo lo necesario en los demás tipos de referencia para apoyar la consecución del objetivo principal.

Se utilizaron las herramientas disponibles en Internet (por ejemplo en el análisis de co-ocurrencias); además se logró obtener un corpus etiquetado verificado manualmente, Clic-TALP

V3.0 de la Universidad Politécnica de Cataluña; el corpus etiquetado facilitó la programación, al depender menos del preprocesamiento, y permitió desarrollar el prototipo inicial; sin embargo, no se podía utilizar con nuevos textos (porque se requiere que estén etiquetados). Para resolver este problema y poder procesar cualquier texto libre se utilizó el corpus Clic-TALP y entrenó el etiquetador TnT logrando salvar esta limitante. Actualmente el sistema en su versión DEMO procesa cualquier texto libre que contenga menos de 4800 unidades léxicas (o tokens) equivalente a 45 KB aprox. Los archivos en formato diferente (html, Word, ps, etc.) deben convertirse a texto puro en ambiente Windows para poder ser procesados.

Es importante mencionar que en este documento no se pretende desarrollar un tratamiento completo de todas las construcciones y fenómenos involucrados en la resolución de referencias y de la anáfora, sino sólo de aquellos que se juzgan más importantes para establecer claramente los conceptos en que se fundamenta la resolución de la anáfora indirecta.

1.7 Organización del documento

Habiendo presentado el panorama del trabajo en la introducción, el capítulo 2 presenta el marco teórico de referencia, o marco conceptual seleccionado y conformado *desde el punto de vista lingüístico*. Se inicia con la descripción de los niveles del lenguaje y el contexto del discurso; se continúa con la descripción del texto y sus propiedades finalizando con el análisis de las referencias del discurso tomando en cuenta la función de los determinantes así como su relación con la elipsis nominal y la anáfora. Después se describen la anáfora directa e indirecta y finalmente se presentan sus características e interrelación por medio de ejemplos comentados que permiten visualizar el modelo computacional a desarrollar para resolver la anáfora directa.

En el capítulo 3, apoyándose en el marco conceptual presentado en el capítulo 2, se describe el método de resolución computacional propuesto y el análisis del sistema requerido. Se desarrolla exponiendo primero como se propone detectar y resolver la anáfora indirecta apoyándose en las expresiones referenciales para finalizar describiendo los algoritmos desarrollados para lograrlo.

En el capítulo 4 se describe la implantación tomando en cuenta el corpus utilizado, la información adicional en diccionarios y desarrollo del prototipo inicial; posteriormente se

presentan los criterios y adecuaciones necesarias, que se hicieron para que el sistema pueda trabajar con texto libre.

El capítulo 5 se comentan: las razones que apoyan el diseño desarrollado desde la selección del corpus más apropiado y del tamaño de la muestra adecuado; la evaluación experimental del método diseñado con un prototipo inicial; la evaluación experimental con archivos con texto libre considerando diferentes condiciones.

El capítulo 6 presenta las conclusiones obtenidas de la investigación realizada; el capítulo 7 describe las aportaciones y las contribuciones logradas, y las tareas pendientes a resolver en el futuro. Finalmente se reportan las referencias citadas y los anexos contienen información adicional que soporta el trabajo realizado y cuya consulta da un panorama más amplio del mismo.