6 Conclusiones

6.1 Resultados obtenidos

El trabajo de investigación realizado permitió desarrollar un método de resolución de la anáfora indirecta que trabaja con un buen nivel de precisión y exactitud, hasta del 92% y 98.51% respectivamente en el prototipo inicial (ver tabla 24), cuando existe información completa disponible en los diccionarios de sinónimos y escenarios. Al adecuarlo para utilizarlo con texto libre la precisión es menor llegando a un 51%, sin embargo el nivel de exactitud se mantuvo en un nivel semejante de 82% comparado con el "experto humano" con un rendimiento global promedio del sistema de 60% de acuerdo con la métrica *F* (ver tabla 33).

Estos indicadores permiten visualizar que el contexto lingüístico basado en el modelo de escenario logra la resolución automática de la anáfora indirecta nominal (meta de este trabajo) con una fuerte dependencia de la información suministrada (compilar automáticamente esta información es un área de oportunidad para trabajos futuros).

Se descubrió la relación existente entre la sintaxis, la semántica y la pragmática, observando que: la sintaxis *sólo marca las expresiones definidas*; apoyándose en la semántica y en la pragmática es posible determinar el tipo de referencia existente (correferencia directa o indirecta) y la relación anafórica indirecta en un texto. Se desarrollo un algoritmo de resolución de correferencias, basado en un diccionario de sinónimos, como requisito previo a la detección de la anáfora indirecta para poder resolverla, por medio de un diccionario de escenarios. Así pues, no se puede hablar de marcadores específicos de cada fenómeno sino de una imbricada red de relaciones que sólo puede resolverse con un algoritmo que de forma integral modele el proceso de lectura del receptor. En otras palabras, el discurso debe verse como un "*conocimiento del receptor que se ve enriquecido paulatinamente con la información recibida y conforme avanza el proceso de lectura*".

6.2 Aportaciones

Las aportaciones específicas de este trabajo son:

- Se descubrió la interrelación existente de los fenómenos de correferencia (directa e indirecta) y la anáfora indirecta y que ambas utilizan expresiones nominales referenciales para manifestar su presencia.
- Se determinó el orden de evaluación requerido para discriminar los fenómenos que sirven como base para detectar la presencia de la anáfora indirecta.
- Se desarrolló un método basado en el modelo de escenario del contexto lingüístico que modela al lector humano, necesario para la resolución de la anáfora indirecta y las correferencias.

6.3 Contribuciones

- Se desarrolló un conjunto de programas que integrados permiten la creación de diccionarios sin repetición de entradas o duplicidad de información logrando así reducir el tamaño de los archivos y los tiempos de acceso a disco; y como consecuencia reducir el tiempo de procesamiento.
- Se desarrolló un conjunto de programas que permiten la extracción de información del diccionario semántico EuroWordNet en Español desde sus archivos de exportación de información.
- Se construyó un diccionario de escenarios extrayendo la información semántica almacenada en el diccionario de EuroWordNet en Español con las relaciones de holonimia, meronimia, y rol necesarias para este trabajo.

De este trabajo surgieron 4 publicaciones, tres son ponencias para congresos internacionales y una es un reporte técnico del estado del arte en anáfora indirecta publicado en el CIC-IPN.

6.4 Recomendaciones y sugerencias para el trabajo futuro

Para mejorar la precisión del sistema desarrollado es necesario perfeccionar la etapa de preprocesamiento, por lo que se requiere:

- Modernizar los etiquetadores actuales con el conocimiento lingüístico (semántico y pragmático) adicional que permita lograr una mayor automatización y precisión. Se espera desarrollar trabajo conjunto a futuro con Brants Thorsten sobre el etiquetador TnT, porque las técnicas utilizadas actualmente han llegado a un límite que obliga a la búsqueda de nuevos caminos para perfeccionarlos.
- Adecuar los etiquetadores actuales con tecnologías de aprendizaje incremental basada en casos y continuar trabajando en conjunto con Montserrat Civit aprovechando las bondades de su trabajo con los corpus etiquetados en Español. Se espera poder trabajar en equipo con el Dr. Aurelio López López del INAOE para apoyar el desarrollo de un etiquetador para el Español.

Para mejorar la base de la información implícita que aumente el poder de resolución del sistema desarrollado es necesario automatizar la construcción de diccionarios, por lo que se requiere:

- Desarrollar una biblioteca de rutinas de procesamiento de cadenas y texto multilingüe que trabaje inicialmente para el Español. Las bibliotecas actuales tienen problemas para el ordenamiento y comparación de cadenas con símbolos o letras específicos del Español (ñ, í, ü, etc.). Esto obligó a revisar y realizar parte del trabajo de elaboración de diccionarios manualmente.
- Involucrarse en el desarrollo de estándares en la construcción de diccionarios semánticos y algoritmos de extracción del significado de textos libres.

Para mejorar el rendimiento global del sistema desarrollado es necesario mejorar los algoritmos de reconocimiento de entidades referenciales, por lo que se requiere:

Investigar a mayor profundidad el reconocimiento de entidades en las expresiones referenciales para poder diferenciarlas cuando se refieren a objetos diferentes del mundo real. Es

necesario ampliar el concepto de identificación para que el sistema pueda manejar objetos similares del mundo real en el mismo texto, por ejemplo:

> El **perro**$_1$ negro de Juan$_2$ mordió al **perro**$_3$ café de Pedro$_4$. El veterinario$_5$ dijo que el **perro**$_3$ debería conservar el vendaje$_6$ por una semana$_7$ para evitar una infección$_8$.

En este párrafo el sistema trabaja bien porque la búsqueda hacia atrás encuentra primero al **perro**$_3$. Pero en el siguiente párrafo fallaría:

> El **perro**$_1$ café de Pedro$_2$ fue mordido por el **perro**$_3$ negro de Juan$_4$. El veterinario$_5$ dijo que el **perro**$_1$ debería conservar el vendaje$_6$ por una semana$_7$ para evitar una infección$_8$.

En esta oración el sistema trabaja mal porque la búsqueda hacia atrás encuentra primero al **perro**$_3$ y el perro mordido es el **perro**$_1$. Es necesario desarrollar un algoritmo que identifique al objeto referido utilizando todos los componentes de la frase nominal para poder evitar una identificación falsa al depender sólo del núcleo de la expresión.

Además es necesario distinguir las frases nominales atributivas de las referenciales y recuperar también la información implícita debida al fenómeno de elipsis. Se está planeando trabajar en conjunto en esta área con Hiram Calvo, aspirante al doctorado en el Laboratorio de Lenguaje Natural del CIC-IPN.