UN MÉTODO PARA LA DESAMBIGUACIÓN SINTÁCTICA DE TIPO COORDINATIVO Y PREPOSICIONAL

A METHOD FOR COORDINATIVE AND PREPOSITIONAL SYNTACTIC DISAMBIGUATION

CARLOS ZAPATA
Grupo de Ingeniería de Software, Escuela de Sistemas, Universidad Nacional de Colombia, czapata@unal.edu.co

KARLA PALOMINO
Grupo de Ingeniería de Software, Escuela de Sistemas, Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín

ROBERTO ROSERO
Grupo de Ingeniería de Software, Escuela de Sistemas, Facultad de Minas, Universidad Nacional de Colombia, Sede Medellín

Recibido para revisar Febrero 17 de 2007, aceptado Mayo 18 de 2007, versión final Mayo 28 de 2007

RESUMEN: El procesamiento del lenguaje Natural (PLN) investiga y formula mecanismos computacionales que permiten la comunicación hombre-máquina. Conceptualmente, un sistema de PLN se divide en tres procesos principales: análisis morfológico, sintáctico y semántico. En cada uno de estos procesos es factible que se presenten múltiples interpretaciones de una misma palabra o frase, según sea el proceso que se esté llevando a cabo; estas interpretaciones dan origen al concepto de ambigüedad. Para resolver la ambigüedad se han propuesto métodos basados en estadística, inteligencia artificial y métodos híbridos, los cuales aún presentan dificultades como el alto consumo de recursos léxicos y computacionales y el uso de elementos pertenecientes a dominios restringidos. En este artículo se propone un método que incluye la definición de un conjunto de reglas heurísticas para desambiguar frases que poseen ambigüedad sintáctica de tipo coordinativo y preposicional. Además, se muestra la implementación del método en el lenguaje python y, combinada con herramientas del paquete NLTK, y se presentan dos casos de estudio para ejemplificar el método.

PALABRAS CLAVE: Procesamiento del Lenguaje Natural, Análisis sintáctico, Información sintáctica y semántica, Ambigüedad sintáctica Coordinativa, Ambigüedad sintáctica preposicional, Desambiguación.

ABSTRACT: Natural Language Processing (NLP) have researched and formulated computational mechanisms to ease Human-Computer Interaction (HCI). From the conceptual point of view, a NLP system can be divided into three main processes: morphology, syntax and semantics. Every process has to deal with multiple interpretations for the same word or phrase; as a result, ambiguity is originated. To solve ambiguity, statistics-based, artificial-intelligence-based, and hybrid methods have been proposed; however, there are still difficulties to be solved, for example wasting of lexical and computational resources and using of restricted-domain elements. Here in this paper we propose a method for solving coordinative and prepositional syntactic ambiguity; this method includes the definition of a set of heuristic rules. Also, we show the implementation of the method using the python language in conjunction with the Natural Language Tool Kit (NLTK), and we exemplify disambiguation of two case studies.

KEYWORDS: Natural Language Processing, Syntactic Analysis, Syntactic and Semantic Information, Coordinative and Preposition Syntactic Ambiguity, Disambiguation.

.

1. INTRODUCCIÓN

El procesamiento del lenguaje natural (PLN) trata los fenómenos lingüísticos de forma Mecanizada mediante sistemas de cómputo [1]. Se define además, como una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos que sean computacionalmente efectivos y que faciliten la interacción hombre-máquina [2]. El PLN surge como una necesidad de automatizar tareas que en la actualidad requieren mucho tiempo para ser realizadas. Conceptualmente, un sistema de PLN divide el análisis de un texto en los siguientes niveles: análisis morfológico, análisis sintáctico y análisis semántico [1], [2], [3]. A su vez, estas tareas comprenden otros procesos, que permiten extraer y evaluar diferentes tipos de información que contribuyen al análisis global de un texto.

En comparación con otros lenguajes, el español es bastante complejo, pues su interpretación se encuentra limitada por la variabilidad de su estructura y la precisión y elaboración de sus reglas de formación [4]. Estas características generan problemas de ambigüedad (diferentes significados para una misma oración), que dependen del tipo de análisis que se esté llevando a cabo. Desde el punto de vista sintáctico, la ambigüedad genera diferentes representaciones para una misma frase [1].

Con el fin de interpretar el lenguaje natural, algunos grupos de investigación han estudiado los diversos tipos de ambigüedad, sus causas y posibles soluciones. Para corregir este problema, se han propuesto, probado e implementado diversas estrategias de desambiguación; algunas de ellas se han basado en métodos estadísticos [5], [7], [8], [9]; otras estrategias se fundamentan en técnicas de inteligencia artificial [10], [11]; también se han expuesto métodos híbridos que han logrado mejores resultados [12].

Los métodos de desambiguación actuales continúan presentando limitaciones para la resolución del problema de la ambigüedad. Los métodos estadísticos consumen muchos recursos tanto computacionales como léxicos; los métodos basados en inteligencia artificial hacen uso de ontologías o redes semánticas que se encuentran restringidas a dominios específicos; los métodos híbridos, en consecuencia, presentan las mismas limitaciones de los métodos estadísticos y de los basados en IA.En general, todos los métodos restringen la sintaxis, el vocabulario y el dominio del texto que se desea analizar [5]; sin embargo, existen algoritmos que ofrecen buenos resultados para un lenguaje restringido, es decir, definido por una gramática de cobertura limitada [5].

En este artículo se propone un método de desambiguación sintáctica para un texto escrito en español. El método se basa en reglas heurísticas que permiten identificar ambigüedades de tipo coordinativo y de tipo preposicional, permitiendo la obtención de los árboles sintácticos más probables para cada una de las frases que hacen parte del texto.

El artículo está organizado así: en la Sección 2 se muestra el marco teórico que fundamenta los conceptos concernientes al lenguaje natural; en la Sección 3 se hace una revisión de los métodos de desambiguación sintáctica que se han desarrollado; en la Sección 4 se propone un método de desambiguación sintáctica para un texto escrito en lenguaje natural que identifica ambigüedades de tipo coordinativo y de tipo preposicional; en la Sección 5 se presenta la aplicación del método, resultados y dificultades; en la Sección 6 se plantean conclusiones acerca del trabajo realizado y el trabajo futuro que se deriva a partir del método propuesto.

2. MARCO TEÓRICO

El procesamiento del lenguaje natural (PLN) ha surgido como una solución a los obstáculos de tipo lingüístico que se generan en la búsqueda de información; en esta búsqueda, el hombre ha optado por automatizar tareas que en la actualidad requieren mucho tiempo para ser realizadas.

El procesamiento del lenguaje natural (PLN) trata todo tipo de fenómenos lingüísticos de forma automática [1], y se define como una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos que sean computacionalmente efectivos y que faciliten la interacción hombre-máquina [2]. De ahí la importancia que ha venido adquiriendo el PLN, ya que no solamente se están solucionando problemas lingüísticos, sino que implícitamente

se está reduciendo el tiempo de procesamiento de la información.

Conceptualmente, un sistema de PLN divide el análisis de un texto en los siguientes niveles: análisis morfológico, análisis sintáctico y análisis semántico [1], que se pueden definir así:

- Análisis morfológico: consiste en determinar la forma, clase o categoría gramatical de cada palabra que hace parte de una oración, haciendo lo que se conoce como etiquetado morfológico.
- Análisis sintáctico: consiste en determinar las funciones de las palabras o grupos de palabras dentro de la oración.
- Análisis semántico: consiste en asignar significados a las estructuras generadas por el analizador sintáctico, es decir se establecen correspondencias entre las estructuras sintácticas y cada palabra dentro de un dominio.

Para la realización de estos análisis, existen herramientas tales como el MPRO [13] y el NLTK (Natural Language Tool Kit) [14], que permiten realizar algunos pasos del proceso de análisis, pero que no poseen módulos de desambiguación; el trazado de los árboles sintácticos de NLTK, por ejemplo, se usará para la propuesta que se presenta en este artículo.
Debido a la variabilidad de la gramática española, a la precisión y elaboración de las reglas de formación del español [6], las tareas de análisis se dificultan, ya que se presentan diferentes tipos de ambigüedad durante cada nivel de análisis, así:

- Ambigüedad Morfológica: Ocurre cuando una palabra que se encuentra en una oración representa más de un rol sintáctico o categoría gramatical dentro de la misma [1], [2], [3].
- Ambigüedad Sintáctica: Se presenta cuando una oración tiene asociada más de una representación sintáctica, es decir, cuando más de una regla gramatical representa dicha oración [1], [2], [3].
- Ambigüedad Semántica: Ocurre cuando una oración posee más de un significado o sentido; se refiere a fenómenos como la homonimia y la polisemia, en los cuales la misma palabra puede tener distintos significados [1], [2], [3], [8].

3. REVISIÓN DEL ESTADO DEL ARTE

La ambigüedad, en el proceso lingüístico, puede presentarse cuando es posible admitir diferentes interpretaciones a partir de la representación de una oración; también, se presenta cuando existe confusión al tener diversas estructuras asociadas a la misma oración. Para desambiguar, es decir, para seleccionar los significados o estructuras más adecuados de un conjunto conocido de representaciones, se requieren diversas estrategias de solución que dependen del tipo de ambigüedad que presente una frase u oración [15].
Con el fin de interpretar automáticamente el lenguaje natural, se han adelantado estudios acerca de los diversos tipos de ambigüedad que puede presentar un texto, sus causas y posibles soluciones. Para corregir el problema de la ambigüedad, se han propuesto, probado e implementado diversas estrategias de desambiguación. Sin embargo, persisten algunas limitaciones en cuanto a los textos que se deseen analizar, ya que se encuentran restringidos a un dominio específico.
El español ha sido uno de los lenguajes más difíciles de tratar, puesto que la mayoría de los recursos léxicos disponibles se encuentran en inglés y para el procesamiento del español se cuenta con recursos muy limitados.
Actualmente existen varios métodos de desambiguación de un texto. Estos métodos se clasifican en: métodos estadísticos, métodos de inteligencia artificial y métodos híbridos.

3.1 Trabajos basados en métodos estadísticos

En [5] se muestra un método supervisado de desambiguación del sentido de las palabras basado en los modelos de Markov (MM) especializados; estos métodos utilizan matrices de probabilidades, donde cada estado corresponde a una categoría morfosintáctica y el

número de estados corresponde al número de categorías asociadas a una palabra. El método consiste en dos tareas fundamentales: la selección de las características relevantes para la tarea de desambiguación, mediante la definición del alfabeto de símbolos utilizado en un MM, y la especialización o redefinición de los estados del modelo a partir de la información disponible en los datos de entrenamiento. Este método requiere recursos computacionales que son escasos para el idioma español, como es el caso de los corpus anotados semánticamente.

En [7] se muestra un método probabilístico, basado en una gramática lexicalizada (gramática que proporciona mayor información sintáctica por cada categoría gramatical). El método combina probabilidades sintácticas, las cuales permiten seleccionar una categoría sintáctica de un conjunto de categorías asociadas a una palabra, y probabilidades semánticas, que posibilitan la selección de una regla sintáctica entre un conjunto de reglas asociadas a una oración.

En [8] se propone un método de aprendizaje supervisado a partir de un corpus de textos anotados semánticamente para la resolución de la ambigüedad semántica de las palabras. Se necesita una fase previa de aprendizaje antes de poder construir y almacenar un clasificador para cada palabra; en esta fase se recogen los ejemplos del corpus y se incorporan al modelo de probabilidad para hacer la estimación de la función de clasificación. El método utiliza los Modelos de Máxima entropía (MME) para realizar la asignación de sentidos a cada palabra y un algoritmo de aprendizaje que permite estudiar los ejemplos y asignar pesos a las palabras que hacen parte de los mismos. Los MME se exponen de una forma más amplia en [16].

En [9] se propone un método de desambiguación léxica. El método consiste en asignar automáticamente el sentido de las palabras que aparecen dentro del contexto de una oración, recurriendo a WordNet Domains [17], el cual se usa para recopilar ejemplos de los diferentes dominios asociados con los significados semánticos de las palabras. El valor agregado de esta propuesta es que etiqueta cada palabra, asignándole los dominios a los cuales puede pertenecer dicha palabra. Los dominios se encuentran ordenados de mayor a menor, de acuerdo con la importancia que tenga la palabra en el dominio.

En general, los métodos estadísticos, aunque resuelven algunos problemas de ambigüedad, consumen muchos recursos léxicos y computacionales, lo cual los hace métodos poco convenientes para el español [18]. Además estos métodos son muy especializados, ya que recurren a fórmulas y estudios complejos que hacen necesaria la presencia de un experto, si se desea mejorar el resultado.

3.2 Trabajos basados en técnicas de inteligencia artificial

En [10] se expone un método de resolución de ambigüedad léxica basado en el Modelo de Espacio Vectorial (MEV). Cada sentido de una palabra es representado con un vector, así como el contexto de la palabra a desambiguar. Las entradas del algoritmo están representadas por los vectores, que son procesados mediante el algoritmo LVQ (Learning Vector Quantization). Mediante una función de similitud se comparan los vectores que representan el contexto de cada palabra a desambiguar con cada uno de los vectores de sus sentidos. El sentido representado por el vector de mayor similitud será el designado como sentido desambiguado.

En [11] se propone un método para la solución de la ambigüedad estructural a partir de suposiciones previas acerca del contexto de la frase. La representación de la frase que cumpla el mayor número de suposiciones, será la elegida. Esta técnica de desambiguación en el lenguaje español, y otras relacionadas con la Inteligencia Artificial, por ejemplo mediante redes neuronales, pueden no resultar apropiadas ni precisas. Por ejemplo, en el entrenamiento de una red neuronal (RN), o en la calibración de los vectores empleados en [10], se requieren repositorios de información muy extensos (redes semánticas, ontologías específicas de un dominio particular, lexicones o corpus), para llegar a una solución coherente; esos repositorios son escasos para el lenguaje español [18], o pueden pertenecer a dominios muy restringidos, lo cual limita la aplicación a esos dominios específicos [18], [19].

3.3 Trabajos con métodos híbridos

En [12] se propone un método de resolución de la ambigüedad estructural haciendo uso de información léxica, sintáctica y semántica. Se combinan tres técnicas que son: reglas ponderadas, patrones de manejo y proximidad semántica. El método contiene módulos que arrojan variantes con pesos y se encargan de recopilar y procesar los pesos arrojados. Estos módulos son:

- El módulo de reglas ponderadas, que trabaja con una gramática independiente del contexto, una gramática computacional y un analizador sintáctico tipo chart.
- El módulo de patrones de manejo, que emplea información léxica de verbos, adjetivos y algunos sustantivos, que obtiene a partir de un corpus del español.
- El módulo de proximidad semántica, que obtiene el grado de proximidad de una palabra o grupo de palabras a partir de una red semántica existente.
- El módulo de votación, que se encarga de recopilar los valores arrojados por cada uno de los módulos explicados anteriormente y elegir la(s) estructura(s) sintáctica(s) correcta(s) según la evaluación de cada módulo.

Al igual que los métodos basados en IA, los métodos híbridos requieren recursos que en general no están disponibles para dominios amplios. Además, como ocurre con los métodos estadísticos, el consumo de recursos computacionales puede llegar a ser alto.

4. PLANTEAMIENTO DEL MÉTODO DE SOLUCIÓN

En este artículo, se propone un método de desambiguación que pretende disminuir el número representaciones sintácticas de una frase que presenta ambigüedad originada en las conjunciones o en las preposiciones. El método se divide en cuatro pasos; el primero de ellos es el análisis sintáctico de la frase, que es realizado mediante el módulo de análisis sintáctico del *Natural Language Tool Kit* (NLTK) [14], un conjunto de herramientas, módulos y tutoriales de procesamiento de lenguaje natural basado en el lenguaje de programación python [20]. En el segundo paso, se procede a determinar el tipo de ambigüedad sintáctica que presenta la frase, ya sea de tipo coordinativo, o de tipo preposicional. El tercer paso es la desambiguación como tal, que depende del tipo de ambigüedad que haya sido detectada. En el cuarto y último paso, se realiza el despliegue de resultados; en este paso, la aplicación implementada durante el proyecto muestra gráficamente el (los) árbol(es) sintáctico(s) con su respectiva frase de origen ya desambiguado(s). El segundo, tercero y cuarto pasos son los aportes específicos de esta propuesta y fueron programados también en el lenguaje python por los autores.

A continuación se detalla cada uno de los pasos que se llevan a cabo para lograr la desambiguación sintáctica de una frase:

- **Análisis sintáctico:**

El análisis sintáctico es realizado mediante la herramienta NLTK, la cual provee dos tipos de algoritmos para este análisis. El primero es un método recursivo; el segundo es un método bottom_up llamado *chart_parser*, que funciona de forma iterativa.

Para el análisis sintáctico se elige el método iterativo, ya que ha demostrado ser más eficiente que los métodos recursivos, en cuanto a tiempo de ejecución [14].

La aplicación se encarga de leer un archivo de texto que contiene un subconjunto de reglas sintácticas que hacen parte de la gramática de contexto libre elegida; este subconjunto está conformado por las estructuras sintácticas más comunes del español. El archivo contiene además un conjunto de palabras con su respectivo rol sintáctico asociado.

- **Determinación del tipo de ambigüedad**

Cuando la frase ha sido analizada sintácticamente, la aplicación, empleando un nuevo módulo programado en el desarrollo de este trabajo, se procede a identificar el tipo de ambigüedad sintáctica que presenta la frase. En el contexto de

este artículo, los tipos de ambigüedad sintáctica pueden ser:

- Ambigüedad Sintáctica Coordinativa: se puede presentar cuando una oración contiene más de una palabra de tipo conjunción. Esta ambigüedad puede ser copulativa, disyuntiva o mixta.

- Ambigüedad Sintáctica Preposicional: se puede presentar, cuando una oración contiene una palabra de tipo preposición.

Cuando se identifica el tipo de ambigüedad, se muestran los árboles sintácticos correspondientes a las estructuras sintácticas que representan la frase y el tipo de ambigüedad que presenta. Sin embargo, es posible que la frase no presente ambigüedad sintáctica; en este caso se despliega una sola representación sintáctica y se notifica que la frase no presenta ambigüedad.

- **Desambiguación**

Luego de identificar el tipo de ambigüedad que presenta la frase, se procede a aplicar las reglas correspondientes a la desambiguación; si la ambigüedad sintáctica es de tipo coordinativo, la aplicación se encarga de identificar el tipo de las conjunciones que hacen parte de la frase, y calcular el nivel de profundidad al que se encuentran dentro de la representación gráfica o árbol sintáctico. Dado el caso de que la ambigüedad identificada sea de tipo preposicional, el sistema se encarga de identificar las preposiciones que conforman la frase y luego procede a consultar en un archivo los sentidos asociados a cada preposición y las palabras con sus roles semánticos asociados.
Ambas estrategias de desambiguación se encuentran definidas por ciertas reglas heurísticas que han sido inferidas, implementadas y aplicadas en diferentes casos de estudio (Véase Sección 4.2) por los integrantes del proyecto.

- **Despliegue de resultados**

Luego de aplicar la estrategia de desambiguación, el sistema muestra gráficamente el (los) árbol(es) sintáctico(s) que según las reglas heurísticas definidas no son ambiguos sintácticamente.

4.1 Alcance del método

Las preposiciones y conjunciones en el español, se clasifican en diferentes grupos. Las preposiciones se clasifican en separables e inseparables y las conjunciones, según la función de correlación que cumplen en la oración, se dividen en coordinantes y subordinantes; dentro del grupo de las conjunciones subordinantes se encuentran otros subgrupos que son: copulativas, disyuntivas, adversativas y alternativas. Finalmente en el grupo de las conjunciones subordinantes se encuentran los subgrupos de conjunciones: causales, comparativas, condicionales, continuativas, ilativas y finales.
El método que aquí se propone identifica las preposiciones separables para realizar el análisis sintáctico de la frase, pero se define un nuevo subconjunto de preposiciones separables posiblemente ambiguas, que permitirán determinar la ambigüedad sintáctica preposicional presente en una frase dada. En la Tabla 1 se muestran las preposiciones separables, y cuáles de ellas son consideradas como posibles preposiciones ambiguas según las reglas definidas por el método.
Sin embargo, la información sintáctica de las palabras que conforman una frase, no es suficiente para llevar a cabo la desambiguación preposicional de la misma; es por ello que se hace necesaria la introducción de información semántica de la preposición que genera ambigüedad y de las palabras que la acompañan en la frase [21].

En la Tabla 1 también se pueden observar los sentidos que representan las diferentes preposiciones y que son aceptados por el método.
Para llevar a cabo la tarea de desambiguación sintáctica coordinativa, el sistema comienza reconociendo las conjunciones que hacen parte de la frase, y que se han definido previamente como conjunciones reconocidas por el sistema (Véanse Tablas 2 y 3). Para la posterior desambiguación, el método define un nuevo grupo de conjunciones conformado por aquéllas que posiblemente generen ambigüedad coordinativa dentro de la frase; estas conjunciones son: Y, O, E, U.
Para llevar a cabo la desambiguación de una frase dada, el método que se propone parte de ciertas suposiciones que son:

- Se cuenta con un corpus etiquetado que se encuentra desambiguado morfológicamente.
- Se cuenta con la información sintáctica de los sustantivos, correspondiente al rol desempeñado en la frase.
- Las frases a desambiguar corresponden con al menos una de las reglas sintácticas que hacen parte de la gramática utilizada.

Tabla 1. Preposiciones Separables posiblemente ambiguas y sentidos aceptados para las preposiciones ambiguas
Table 1. Separable and possibly ambiguous prepositions and accepted senses for them

Preposiciones Separables	Preposiciones Sintacticamente ambiguas	Sentido
A	X	Lugar Tiempo Instrumento
Ante		
Bajo		
Cabe		
Con	X	Contenido Compañía Instrumento
Contra		
De	X	Materia Pertenencia Origen
Desde		
En	X	Lugar Tiempo
Entre		
Hacia		
Hasta		
Para		
Por		
Según		
Sobre		
Tras		

Tabla2. Conjunciones Coordinantes
Table 2. Coordinative conjuctions

Conjunciones coordinantes		
Copulativas	**Disyuntivas**	**Adversativas**
Y	O	Aunque
E	U	Pero
Ni	Sea	Mas
Que	Bien	Empero
		Sino
		Siquiera

Tabla 3. Conjunciones Subordinantes
Table 3. Subordinative Conjunctions

Conjunciones subordinantes		
Causales	**Comparativas**	**Condicionales**
Pues	Como	Si
Porque		
Conjunciones subordinantes		
Finales	**Ilativas**	**Temporales**
Para	Aunque	Cuando
Porque	Luego	Antes
	Pues	Luego
		Después

Estas suposiciones, permiten desarrollar el método, pero no es necesaria la existencia de estos requisitos para que sea posible la aplicación del método de desambiguación.

4.2 Reglas Heurísticas

El método está basado en reglas heurísticas, las cuales fueron identificadas e inferidas por los autores después de un proceso de análisis de múltiples frases que presentaban los tipos anotados de ambigüedad. Estas reglas heurísticas han sido divididas en tres grupos, para su mayor comprensión y facilidad en la implementación:

- **Reglas de identificación de la ambigüedad:**

Estas reglas permiten determinar qué tipo de ambigüedad sintáctica presenta una frase.

Regla 1: Si una frase contiene más de una conjunción sintácticamente ambigua y dichas conjunciones pertenecen al grupo de conjunciones coordinantes copulativas entonces la frase presenta ambigüedad coordinativa copulativa.

En la frase: *María* ***y*** *Pedro* ***y*** *David estudian inglés* se presenta este tipo de ambigüedad, ya que se identifican dos conjunciones copulativas consideradas por el sistema como posiblemente ambiguas.

Regla 2: Si una frase contiene más de una conjunción sintácticamente ambigua y dichas conjunciones pertenecen al grupo de conjunciones coordinantes disyuntivas, entonces la frase presenta *ambigüedad coordinativa disyuntiva.*

En el caso: *María o Pedro o David estudiarán inglés* se identifican dos conjunciones disyuntivas consideradas posiblemente ambiguas, por lo tanto según la regla, la frase presenta ambigüedad coordinativa disyuntiva.

Regla 3: Si una frase contiene más de una conjunción sintácticamente ambigua y dichas conjunciones pertenecen al grupo de conjunciones coordinantes disyuntivas o al grupo de conjunciones coordinantes copulativas entonces la frase presenta *ambigüedad coordinativa mixta.*

En la frase: *María y Pedro o David estudiarán inglés* se identifica una conjunción disyuntiva y otra copulativa, que indican que la frase presenta una ambigüedad coordinativa mixta.

Regla 4: Si una frase contiene al menos una preposición separable, que sea sintácticamente ambigua entonces la frase presenta *ambigüedad preposicional*.

La frase: *Juan va a la fiesta con la novia* contiene dos preposiciones: la primera de ellas es la preposición "a" y la segunda es la preposición "con"; estas preposiciones son agrupadas por el sistema como separables y posiblemente ambiguas.

- **Reglas de extracción de información semántica:**

Estas reglas permiten reunir la información semántica necesaria tanto de la(s) preposición(es) que genera(n) ambigüedad como de las palabras que la(s) acompañan; esta información semántica se conoce como roles semánticos de una palabra [21].

Si la Regla 4 se cumple, entonces:

Regla 5: Se identifica la preposición que ha generado la ambigüedad sintáctica.

Regla 5.1 La preposición es "a":

Regla 5.1.1 Si la preposición se encuentra sucedida por un sintagma nominal cuyo núcleo sea un sustantivo que puede representar locación, entonces el sentido de la preposición es de *Lugar*.
Regla 5.1.2 Si la preposición se encuentra sucedida por un sintagma nominal cuyo núcleo sea un sustantivo que actúe o represente un punto en el tiempo, entonces la preposición es de *tiempo.*
Regla 5.1.3 Si la preposición se encuentra sucedida por un sintagma nominal cuyo núcleo sea un sustantivo que actúe o represente un medio o instrumento, entonces la preposición es de *Instrumento.*

Regla 5.2 La preposición es "con":

Regla 5.2.1 Si la preposición se encuentra sucedida por un sustantivo que puede representar instrumento entonces la preposición es de *instrumento*
Regla 5.2.2 Si la preposición se encuentra sucedida por un sustantivo que actúe o represente una persona u objeto animado entonces la preposición es de *compañía.*
Regla 5.2.3 Si la preposición se encuentra sucedida por un sustantivo que actúe o represente una sustancia o material entonces la preposición es de *contenido.*

Regla 5.3 La preposición es "de":

Regla 5.3.1 Si la preposición se encuentra sucedida por un sustantivo que puede representar un tipo de material entonces la preposición es de *materia.*
Regla 5.3.2 Si la preposición se encuentra sucedida por un sustantivo que actúe o represente una persona o en general un objeto animado entonces la preposición es de *pertenencia*.
Regla 5.3.3 Si la preposición se encuentra sucedida por un sustantivo que actúe o represente un lugar entonces la preposición es de *origen*.
Regla 5.4 La preposición es "en":
Regla 5.4.1 Si la preposición se encuentra sucedida por un sustantivo que puede representar un lugar entonces la preposición es de *lugar.*
Regla 5.4.2 Si la preposición se encuentra sucedida por un sustantivo que represente un punto en el tiempo entonces la preposición es de *tiempo*

- **Reglas de desambiguación:**

Estas reglas permiten aplicar una estrategia de desambiguación posterior al cumplimiento de alguna de las reglas de identificación de la ambigüedad.
Si se cumple la Regla 1 entonces:

Regla 6: Para cada uno de los árboles sintácticos

generados en el análisis sintáctico, se determina el nivel de profundidad, al cual se encuentra cada una de las conjunciones presentes en la frase.

Regla 6.1 Para cada árbol sintáctico se suman los niveles de profundidad hallados en la Regla 6 correspondientes a las conjunciones copulativas.
Regla 6.2 El árbol sintáctico elegido es aquel cuya suma de niveles de profundidad halladas en la Regla 6.1 sea la mayor. Si se presentan empates entre las sumas de niveles de profundidad de varios árboles sintácticos, se presentan todos los árboles, lo cual implica que sólo se pudo aplicar la desambiguación hasta ese resultado.

Si se cumple la Regla 2 entonces:

Regla 7: Para cada uno de los árboles sintácticos generados en el análisis sintáctico, se determina el nivel de profundidad, al cual se encuentra cada una de las conjunciones presentes en la frase.
Regla 7.1 Para cada árbol sintáctico se suman los niveles de profundidad hallados en la Regla 6 correspondientes a las conjunciones disyuntivas.
Regla 7.2 El árbol sintáctico elegido es aquel cuya suma de niveles de profundidad halladas en la Regla 7.1 sea la mayor. Si se presentan empates entre las sumas de niveles de profundidad de varios árboles sintácticos, se elige únicamente el primero de ellos.

Si se cumple la Regla 3 entonces:

Regla 8: Para cada uno de los árboles sintácticos generados en el análisis sintáctico, se determina el nivel de profundidad, al cual se encuentra cada una de las conjunciones presentes en la frase.
Regla 8.1 Por cada árbol sintáctico se comparan los niveles de profundidad hallados en R8 correspondientes a las conjunciones copulativas que hayan sido reconocidas.
Regla 8.2 El árbol sintáctico elegido es aquel cuyo nivel de profundidad correspondiente a la(s) conjunciones copulativas sea el mayor.

Si se presentan empates en la Regla 8.1, se eligen los árboles sintácticos que hayan generados empates y que hayan obtenido el mayor nivel de profundidad.
Si se cumple la Regla 5.1.1 entonces:

Regla 9: Se elige la representación cuyo nivel de profundidad para la preposición "con" sea la menor.

Si se cumple la Regla 5.1.3 entonces:

Regla 10: Se elige la representación sintáctica cuyo número de niveles sea el mayor.

Si se cumple la Regla 5.2.2 entonces:

Regla 11: Se calcula el nivel de profundidad de la preposición "con" para cada una de las representaciones gráficas.

Si se cumple la Regla 5.2.3 entonces:

Regla 12: Se calcula el nivel de profundidad de la preposición "con" para cada una de las representaciones sintácticas.

Si se cumple la Regla 5.3.1 entonces:

Regla 13: Se elige la representación sintáctica, donde el nivel de profundidad de la preposición "de" sea el menor.

Si se cumple la Regla 5.3.2 entonces:

Regla 14: Se elige la representación sintáctica, donde el nivel de profundidad calculado de la preposición "de" sea el menor.

Si se cumple la Regla 5.4.1 entonces:
Regla 15: Se elige la representación sintáctica para el cual el nivel de profundidad en el que se encuentra la preposición "de" es la mayor.
En general, si en cualquiera de los casos se presentan empates en el nivel de profundidad al que se encuentran las preposiciones, se elijen las representaciones que presentan empates.

5. APLICACIÓN DEL MÉTODO Y RESULTADOS OBTENIDOS

Como una forma de validación de la efectividad de las reglas, el método se programó en lenguaje python, como un módulo complementario a la herramienta NLTK, la cual se emplea para la realización del análisis sintáctico (en esencia, la construcción de los árboles sintácticos). El núcleo del proceso de desambiguación se basa en las reglas presentadas en la Sección anterior.

La primera interfaz que presenta la aplicación, permite ingresar una frase para ser posteriormente analizada y desambiguada (Véase Figura 1).
En esta interfaz, el usuario escribe la frase que desea desambiguar y da clic en el botón ingresar; internamente, el sistema lleva a cabo el análisis sintáctico de la frase y muestra los árboles sintácticos correspondientes a la frase ingresada (Véase Figura 2).
Luego, el sistema muestra un mensaje al usuario que le informa el tipo de ambigüedad sintáctica que presenta la frase (Véase Figura 3). Para entregar este mensaje, la aplicación internamente se encuentra validando las Reglas 1 a 4 que hace parte de las reglas de identificación de la ambigüedad (Véase Sección 4.2).
Posteriormente, el sistema se encarga de aplicar las reglas de desambiguación y finalmente mostrar el (los) árbol(es) sintáctico(s) ya desambiguados (Véase Figura 4). Además, se despliega una ventana que indica las reglas que fueron aplicadas en el proceso.

Figura 1. Interfaz para Ingresar Frase
Figure 1. Snapshot of the input-phrase interface

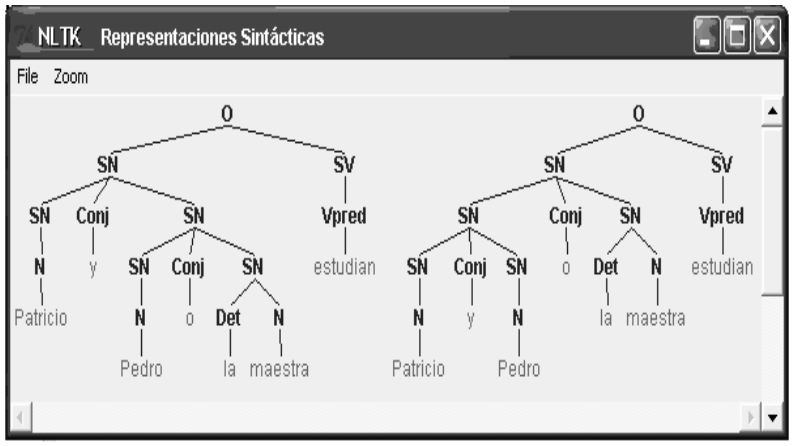

Figura 2. Resultados del Análisis Sintáctico
Figure 2. Results of the Syntactic Analysis

Figura 3. Información del Tipo de ambigüedad
Figure 3. Information about the ambiguity type

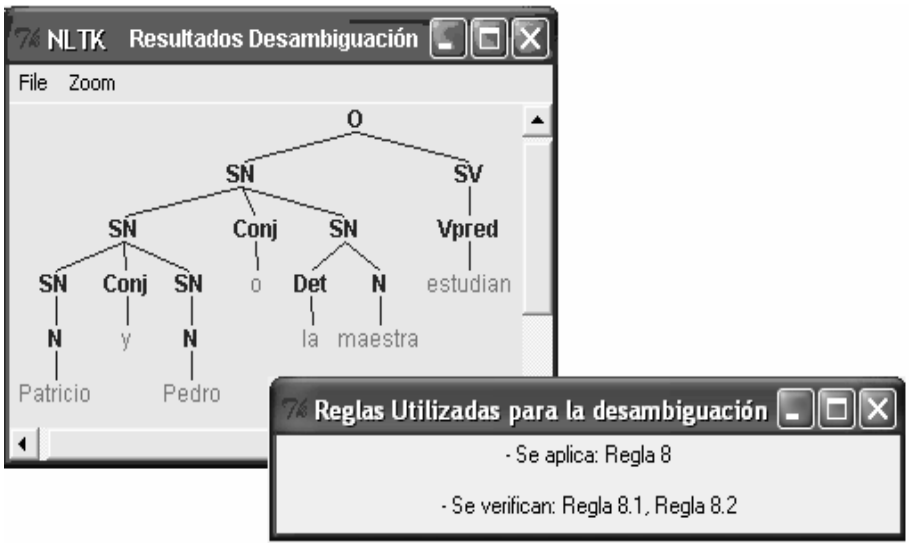

Figura 4. Resultados de la desambiguación
Figure 4. Results of the Disambiguation process

A continuación se presentan dos casos de estudio; el primero de ellos muestra una frase que presenta ambigüedad sintáctica coordinativa, mientras que el segundo está caracterizado por una frase con ambigüedad sintáctica preposicional. En ambos casos se aplica el método de desambiguación y se muestran los resultados obtenidos durante el proceso.

CASO 1: *"Patricio y Pedro o la maestra estudian"*
Inicialmente se escribe la frase que se desea desambiguar. Luego se da clic en el botón Ingresar (Véase Figura 1)
El análisis sintáctico de la frase genera dos representaciones sintácticas gráficas. Cada una de estas representaciones corresponde a diferentes agrupaciones de las categorías gramaticales. En la primera representación gráfica (Véase Figura 2), se puede observar que la conjunción "Y" afecta a Patricio y al grupo nominal conformado por *Pedro o la maestra*. La conjunción "O" afecta a *Pedro* y

la maestra. Esto se puede apreciar por la forma como se encuentran agrupados los Sintagmas nominales (SNs).

En la segunda representación (Véase Figura 2) sucede que la conjunción "Y" afecta a *Patricio* y a *Pedro*. La conjunción "O" afecta al grupo nominal conformado por *Patricio* y *Pedro* y al sintagma nominal representado por *la maestra*.
Tras realizar el análisis sintáctico de la frase, el sistema identifica que la frase presenta ambigüedad sintáctica coordinativa mixta puesto que "Y" es una Conjunción copulativa, y "O" una Conjunción disyuntiva, entonces se cumple la Regla 3; al cumplirse esta regla, debe llevarse a cabo la verificación de las Reglas 8, 8.1 y 8.2. En este caso, los resultados favorecen a la representación grafica 2 (Véase Tabla 4), puesto que el nivel de profundidad para la conjunción "Y" en la representación 2 es mayor que el nivel de profundidad para la conjunción "Y" en la representación 1 (Véase Tabla 5).

Tabla 4. Resultados de los niveles de profundidad de cada representación sintáctica
Table 4. level depth results for every syntactic representation

Conjunción	Representación 1	Representación 2
Y	3	2
O	2	3

Como en este caso sólo se encuentra una ocurrencia por conjunción, únicamente se compara el nivel de profundidad de la conjunción "Y" calculado para cada árbol.

En la Figura 4 se presenta el árbol que ha sido elegido por el sistema mediante el método de

desambiguación sintáctica para la ambigüedad sintáctica coordinativa mixta del caso de estudio.

Tabla 5. Elección de la representación sintáctica
Table 5. Selection of the syntactic representation

Número Representación	Nivel de Profundidad	Representación elegida
1	2	
2	3	**X**

CASO 2: *"Patricio va a la playa con la novia"*
Los resultados obtenidos tras realizar el análisis sintáctico de la frase, muestran dos representaciones sintácticas de la misma. Se identifica una posible ambigüedad sintáctica de tipo preposicional, puesto que la frase contiene dos preposiciones posiblemente ambiguas, que son: "a" y "con". El sistema, que emplea el método propuesto, indica mediante un mensaje cuál regla heurística se cumple para llegar a concluir qué tipo de ambigüedad posee la frase (Véase la Figura 5).

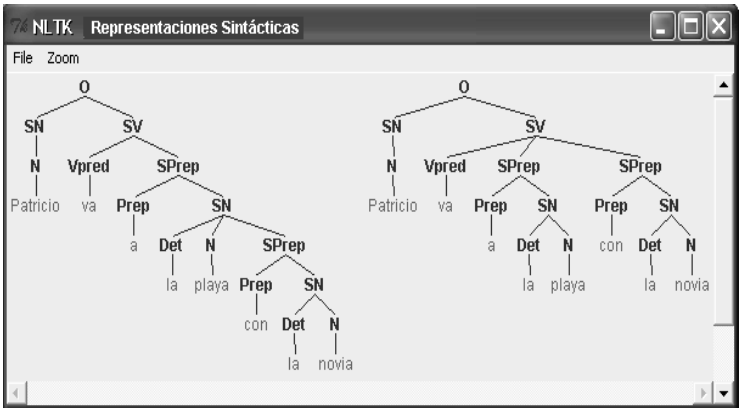

Figura 5 Resultados del Análisis Sintáctico
Figure 5. Results of the Syntactic Analysis

De igual forma, el sistema procede a mostrar un mensaje que señala cuáles reglas fueron utilizadas para resolver la ambigüedad de la frase ingresada (Véase Figura 6)

Figura 6 Información del Tipo de ambigüedad
Figure 6. Information about the ambiguity type

Para realizar la desambiguación correspondiente, el sistema aplica la regla de desambiguación 5.2, y Verifica las reglas 5.2.1, 5.2.2 y 5.2.3. En este caso, los resultados obtenidos tras la aplicación de dichas reglas fueron:
Para el caso de la preposición "con", indica Compañía ya que se encuentra sucedida por el Sintagma nominal cuyo núcleo está representado por el sustantivo *novia* que representa una persona, que corresponde a la regla 5.2.2. (Véase Sección 4.2). Luego de identificados la preposición y su posible sentido, se procede a desambiguar la frase mediante la Regla 10, en este caso se obtienen los resultados que se muestran en la Tabla 6.

Tabla 6. Elección de la representación sintáctica
Table 6. Selection of the syntactic representation

Número Representación	Nivel de Profundidad	Representación elegida
1	6	**X**
2	4	

De la Tabla 6 se infiere que la representación sintáctica elegida según las reglas heurísticas es la representación 1 ya que el nivel de profundidad de la preposición "con" calculado, es mayor que el nivel calculado para la representación 2 (Véase Figura 7).

Figura 7. Resultados de la desambiguación
Figure 7. Results of the Disambiguation process

6. CONCLUSIONES Y TRABAJO FUTURO

-En este artículo se presentó un método para la desambiguación de tipo coordinativo y preposicional. Se definieron las reglas heurísticas que permiten definir el tipo de ambigüedad que presenta la frase, para posteriormente tratar la ambigüedad y presentar una o varias estructuras que puedan contribuir a precisar el sentido de las frases, pero tomando en cuenta únicamente el análisis morfológico y sintáctico.
-Una ventaja del método propuesto, es que suministra información concerniente al tipo de ambigüedad sintáctica de una frase; esta información puede ser reutilizada en tareas posteriores del procesamiento del lenguaje natural, que a su vez permitan automatizar diversas actividades que involucren este tipo de procesos.
El método propuesto hace uso de la herramienta NLTK [14], la cual ha mostrado buenos resultados en el campo del PLN, y la complementa con Código desarrollado en el lenguaje de programación python [20], que se caracteriza por su flexibilidad y facilidad de programación. Estas herramientas presentan ventajas para el futuro mejoramiento y actualización del sistema.
-La desambiguación y análisis sintácticos de una frase se logran mediante la conjugación de diferentes clases de información lingüística. El análisis sintáctico de una frase requiere de información morfológica y la desambiguación requiere de información sintáctica y semántica; es por ello que el análisis sintáctico es considerado una de las tareas más complejas y completas que hacen parte del PLN.
-En relación con los métodos empleados en la literatura, el sistema que implementa esta propuesta no posee un alto consumo de recursos computacionales o léxicos (sólo se requiere un lexicón muy sencillo). Además, no se necesitan ontologías del dominio ni corpus específicos de ciertos temas, lo que le suministra generalidad para el trabajo de desambiguación.
-Gracias al lenguaje de programación python y al NLTK es posible la integración tanto de una interfaz Web como de una base de datos en Oracle® que permitan lograr una mayor accesibilidad y robustez del sistema. Este es considerado uno de los trabajos que pueden dar continuidad a esta propuesta. Igualmente, la implementación de reglas heurísticas para otros tipos de ambigüedad se podrían considerar como extensiones a este trabajo.

REFERENCIAS

[1] MOONEY, RAYMOND J. Fundamentals, Parte I caps. II, III, IV, V. Oxford Handbook of Computational Linguistics, Oxford University Press. (Ruslan Mitkov Ed.). 2003.

[2] ALLEN, J. Natural language understanding. California: The Benjamin/Cummings Publishing Company. 1987.

[3] HAUSSER, R. Foundations of computational linguistics*:* human_computer comunication in natural language, Berlin: Springer. 2001.

[4] MORENO, L., PALOMAR, M., MOLINA, A., y FERRÁNDEZ, A. Introducción al Procesamiento del Lenguaje Natural. (Ed. Servicio de Publicaciones Universidad de Alicante). Universidad de Alicante. 1999.

[5] MOLINA, A. Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático [PhD tesis]. Universidad Politécnica de Valencia Valencia, 2004.

[6] ZAPATA, C., ARANGO, F. Los modelos verbales en lenguaje natural y su utilización en la elaboración de esquemas conceptuales para el desarrollo de Software: Una revisión crítica. Revista Universidad EAFIT. Vol. 41. Pp 77-95. 2005.

[7] MIYAO, Y., TSUJII J. A model of syntactic disambiguation based on lexicalized grammars. Memorias La séptima conferencia sobre aprendizaje de Lenguaje natural. Edmonton, Canada. Mayo, 2003.

[8] SUÁREZ, CUETO A. Resolución de la ambigüedad semántica de las palabras mediante modelos de probabilidad de máxima entropía [PhD Tesis]. Universidad de Alicante. Alicante, 2004.

[9] VÁZQUEZ S., MONTOYO A., RIGAU G. Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes. Procesamiento del Lenguaje Natural, Revista nº 31. Pp 141-149. 2003.

[10] MARTÍN, VALDIVIA M. TERESA, GARCÍA, VEGA M., UREÑA, LÓPEZ L. ALFONSO. Resolución de la ambigüedad mediante redes neuronales. Procesamiento del Lenguaje Natural, Revista nº 29. Pp 39-45. 2002.

[11] KNOTT, A. AND VLUGTER, P. Syntactic disambiguation using presupposition resolution in Proceedings of the 4th Australasian Language Technology workshop. Melbourne. 2003.

[12] GALICIA-HARO, S., GELBUKH, A. y Bolshakov, Igor A. Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes. Procesamiento del Lenguaje Natural, Revista nº 27. Pp 55-63. 2001.

[13] HALLER, J., DONOSO, A., RAMIREZ, Y. MPRO un programa para el análisis morfológico y sintáctico de textos en español. Procesamiento del Lenguaje Natural, Revista nº 29. pp. 307-308. 2002.

[14] Natural Language Toolkit. http://nltk.sourceforge.net/ [Citado 22 de Noviembre de 2006].

[15] GALICIA, HARO S. Análisis sintáctico conducido por un diccionario de patrones de manejo sintáctico para lenguaje español [PhD]. Instituto Politécnico Nacional. Mexico D.F. 2000.

[16] JAYNES, E.T. (Notes on present status and future prospects), en W.T. Grandy y L.H. Schick, editores, Maximum Entropy and Bayesian Methods. Pp. 1-13. (1990).

[17] MAGNINI, BERNARDO Y C. Strapparava (Experiments in Word Domain Disambiguation for Parallel Texts), en Proceedings of the ACL Workshop on Word Senses and Multilinguality, Hong Kong, China. 2000.

[18] PEREZ M. PASCQA: Búsqueda de Respuestas con base en anotación predictiva de contextos léxico-sintácticos [PhD tesis]. Instituto Nacional de Astrofísica, Óptica y Electrónica Sta. Ma. Tonantzintla, Pue. 2006.

[19] CARRERRO F., GOMEZ J., DE BUENAZA M., MATA J. y MAÑA M. Acceso a la información bilingüe utilizando ontologías específicas del dominio biomédico. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, pendiente publicación.

[20] The Python Programming Language. http://www.python.org/. [Citado 22 de Noviembre de 2006].

[21] NAVARRO B., MOREDA, P., FERNÁNDEZ, B. et al. Anotación de roles semánticos en el corpus 3LB. IX Ibero-American Conference on Artificial Intelligence. 2004.