Una aproximación para resolución de ambigüedad estructural empleando tres mecanismos diferentes*

Sofía N. Galicia-Haro, **Alexander Gelbukh** e **Igor A. Bolshakov**

Centro de Investigación en Computación
Instituto Politécnico Nacional
Av. Juan de Dios Bátiz, 07738 México, D. F.
+52 57296000 ext. 56544, fax +52 5586-2936
{sofia, gelbukh, igor}@cic.ipn.mx

Resumen La ambigüedad estructural es uno de los problemas más difíciles de resolver en sistemas de procesamiento de lenguaje natural. Consideramos dos tipos de resolución de ambigüedad estructural que pueden emplearse en el análisis de textos sin restricciones: conocimiento léxico y cierta clase de contexto. En este trabajo, proponemos un modelo basado en tres diferentes mecanismos para revelar la estructura sintáctica correcta y un módulo de clasificación para obtener las estructuras más probables para la oración analizada.

Nuestro modelo está dirigido al análisis de textos sin restricciones y las herramientas desarrolladas no requieren ninguna desambiguación de marcas morfológicas ni ningún tipo de marcas sintácticas.

1 Introducción

La ambigüedad estructural es uno de los problemas más difíciles de resolver en sistemas de procesamiento de lenguaje natural. La ambigüedad estructural se da porque la sola información sintáctica no es suficiente para realizar una decisión única de asignación de estructura. Investigaciones recientes han desarrollado las gramáticas independientes del contexto probabilísticas, como un medio para seleccionar entre análisis sintácticos alternativos de la misma cadena de palabras, es decir, para la desambiguación.

Los resultados desalentadores pueden explicarse porque esas gramáticas son incapaces de expresar las dependencias entre palabras. Además de que para obtener una cobertura alta en análisis sintáctico preciso se requiere información léxica detallada [Magerman, 95; Collins, 96; Charniak, 97].

Líneas actuales de investigación introducen dependencias léxicas. Por ejemplo, empleando estadísticas de grupos nominales básicos y de pares de palabras [Collins, 99], empleando atracción léxica entre palabras de contenido [Yuret, 98].

Sin embargo, en lenguajes con mayor empleo de preposiciones simples y compuestas, los grupos nominales incluyen grupos preposicionales lo que incrementa la ambigüedad de enlaces de grupos. Así mismo, las estadísticas de pares de palabras tienen un mayor impacto en el análisis sintáctico de lenguajes con mayor restricción en el orden de palabras. En lenguajes como el español, con menores restricciones en el orden de palabras y mayor empleo de preposiciones, la obtención de las estructuras de argumentos para palabras permite reducir las variantes de enlaces de grupos nominales y preposicionales, por lo que se requiere representar o aprender estructuras de argumentos de palabras, estableciendo sus dependencias con preposiciones.

En nuestra opinión, la desambiguación estructural requiere la incorporación de conocimiento semántico basado en cierta clase de contexto local, además de la incorporación de conocimiento léxico basada en dependencias entre palabras. Pero la incorporación de todo ese conocimiento implica un enorme trabajo de codificación manual de información, que restringe su aplicación a dominios limitados. Aún incluyendo conocimiento léxico y semántico en un analizador sintáctico, dada la imposibilidad de detallarlo completamente, se obtendría una cierta cantidad de estructuras, por lo que se requeriría un mecanismo de desambiguación.

En este trabajo, proponemos un modelo para análisis sintáctico y desambiguación basado en dependencias léxicas entre palabras predicativas

* Trabajo hecho con apoyo parcial del CONACyT, SNI y CGEPI-IPN, México.

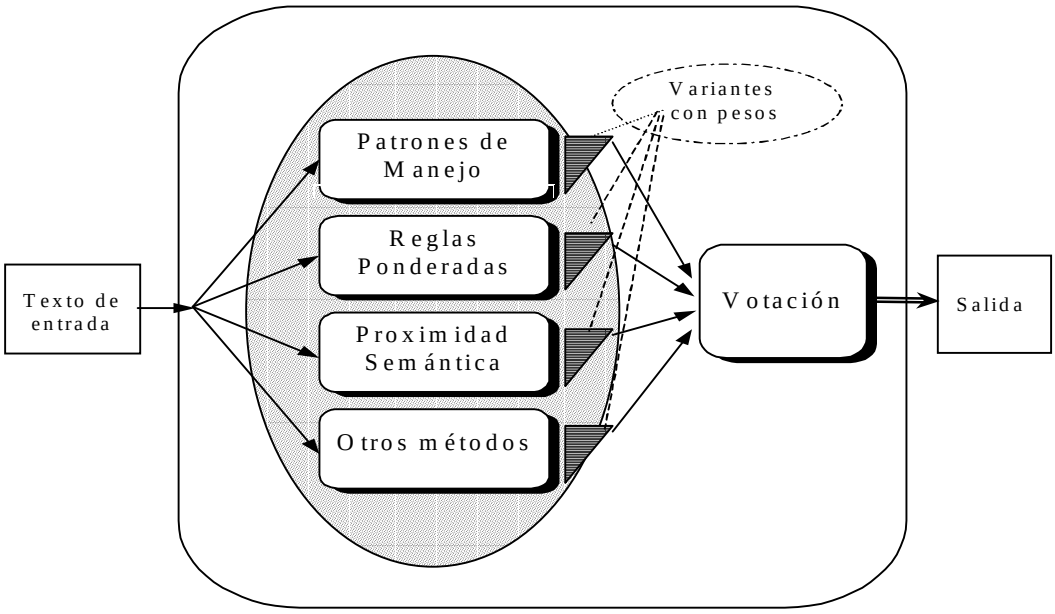

Figura 1. Modelo para asignación de estructuras sintácticas y desambiguación.

y en proximidad semántica. El primero sigue el enfoque de dependencias mediante patrones de manejo sintáctico basados en los esquemas de rección descritos en la Teoría Sentido ⇔ Texto. La proximidad semántica se sustenta en el empleo de una red semántica para incorporar contexto local en un analizador que sigue el enfoque de constituyentes mediante una gramática independiente del contexto (CFG). Para la desambiguación sintáctica proponemos la clasificación de la totalidad de variantes de las estructuras obtenidas, por medio de un peso asignado.

Nuestro modelo está dirigido al análisis de textos sin restricciones y para compilar los recursos necesarios desarrollamos herramientas que no requieren ninguna desambiguación de marcas morfológicas ni ningún tipo de marcas sintácticas.

Primero presentamos el modelo general y enseguida cada uno de los componentes.

2 Modelo general

El modelo asigna estructuras sintácticas en el análisis de textos sin restricciones mediante tres módulos diferentes que operan en paralelo. Cada módulo corresponde a cada uno de los conocimientos mencionados: léxico, sintáctico y semántico, es decir, cada uno se construye basándose en un método diferente que representa un conocimiento específico. Cada módulo proporciona un conjunto de variantes con pesos basados en características satisfechas de cada método. Así que la salida de cada módulo da una medida cuantitativa de la probabilidad de cada estructura sintáctica. Mediante esta medida, un módulo de votación clasifica las variantes para que en el tope aparezcan las más probables de ser las correctas.

El modelo general se presenta en la Figura 1. Las variantes de cada grupo, con mayores probabilidades, constituyen la entrada al módulo de votación, donde se seleccionan las más adecuadas. Dado el carácter cuantitativo del modelo, en el futuro puede incluir otros métodos. Los tres mecanismos considerados actualmente son:

- Reglas ponderadas.
- Patrones de manejo
- Proximidad semántica

El modelo requiere entonces la compilación de tres diccionarios: el de patrones de manejo, la red semántica y las reglas de la CFG extendida.

3 Reglas ponderadas

Es uno de los modelos de resolución de ambigüedad sintáctica más simple pero mucho más cómodo para aplicar y para compilar los recursos necesarios. Para este módulo creamos una gramática independiente del contexto para

el español, una gramática computacional, y un analizador sintáctico tipo *chart*.

La gramática que necesitamos en este caso, dado que no es el método más importante, no requiere condiciones óptimas en cuanto a cobertura y precisión. Nuestra gramática pretende considerar las construcciones más comunes, e incluye las siguientes mejoras:

- Restricción de concordancia, en género, número y persona.
- Inclusión del elemento rector.
- Inclusión de relaciones sintácticas
- Inclusión de elementos de puntuación.
- Inclusión de marca semántica de tiempo.
- Pesos estadísticos.

El elemento rector se requiere para la transformación de las estructuras de constituyentes a estructuras de dependencias, los pesos estadísticos se emplean para graduar el número de reglas que se usan en el análisis.

La gramática que creamos se apoya en las marcas morfológicas que contienen las palabras del corpus LEXESP[1]. Este corpus no contiene desambiguación de POS por lo que el número de análisis es mayor. Esta aparente desventaja tiene su contraparte, si el desambiguador de POS no es de muy buena calidad entonces ocasionará que no se realice el análisis sintáctico de algunas oraciones o que de antemano se orille a un análisis sintáctico incorrecto. El corpus tiene las categorías PAROLE [Civit & Castellón, 98].

La información sola de categorías de POS no nos ayuda a asignar pesos que diferencien las variantes correctas. Emplear las reglas para diferenciar grupos implica el uso de métodos complejos para hacer una clasificación de árboles basándose en la cuál se podrían asignar valores cuantitativos. El peso de las reglas se utiliza directamente en el método por lo que siempre se obtienen las variantes con menor peso en general, es decir, con mayor prioridad. Solamente cuando se utilizan prioridades menores se utilizan reglas con diferentes prioridades.

El análisis de la labor requerida para realizar la clasificación y la asignación de valores, comparada contra los resultados de un método que no distingue información léxica y da estructuras iguales por categorías gramaticales nos hizo proponer una asignación de pesos por igual para todas las variantes, con la finalidad de que los métodos de PMA y de proximidad semántica sean los que hagan emerger las variantes correctas.

[1] El corpus LEXESP nos fue proporcionado amablemente por H. Rodríguez de la Universidad Politécnica de Cataluña, en Barcelona, España.

4 Patrones de manejo sintáctico

Este método se basa en conocimiento lingüístico que adquieren los hablantes nativos durante el aprendizaje de su lenguaje, por lo que se considera el método principal. Este método es el más práctico para solucionar la mayoría de los problemas de ambigüedad. Aunque por sí mismo, este método no es suficiente para el análisis sintáctico de textos sin restricciones, por lo que se consideraron los otros modelos. El conocimiento descrito en este módulo es la información léxica de verbos, adjetivos y algunos sustantivos del español, para enlazar las frases que realizan las valencias. No es posible establecer ese conocimiento mediante reglas o algoritmos pero es posible obtener la información léxica a partir de un corpus.

El método se basa en la teoría Sentido ⇔ Texto (*Meaning ⇔ Text Theory*, MTT) [Mel'cuk, 88], donde con la ayuda de una tabla de Esquemas de Rección (*Government Patterns*, GP) [Steele, 90], se relacionan los participantes semánticos o actantes con los complementos de la cabecera del artículo lexicográfico, es decir, la información de correspondencia entre las valencias semánticas y sintácticas de la cabecera del artículo lexicográfico.

Los GP describen también todas las formas en que se realizan las valencias sintácticas y la indicación de obligatoriedad de la presencia de cada actante, si es necesario.

Después de la tabla de GP se presentan dos secciones: restricciones y ejemplos. Las restricciones consideradas en los GP son de todo tipo: semánticas, sintácticas o morfológicas. La sección de ejemplos cubre todas las posibilidades: ejemplos para cada actante, ejemplos de todas las posibles combinaciones de actantes y finalmente los ejemplos de combinaciones imposibles o indeseables.

La parte principal de la tabla de GP es la lista de valencias sintácticas de la cabecera del artículo lexicográfico. Se listan de una manera arbitraria pero se prefiere el orden de incremento en la oblicuidad: sujeto, objeto

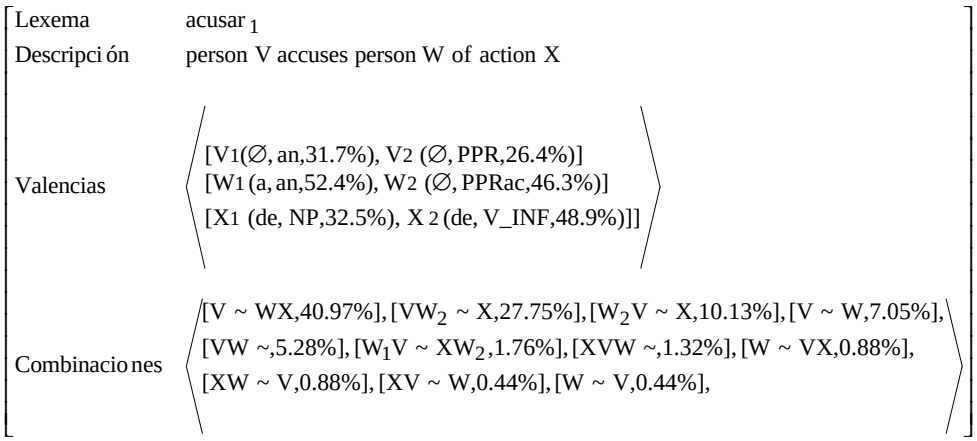

Figura 2. Patrón de manejo sintáctico avanzado para el verbo *acusar*$_1$

directo, objeto indirecto, etc. También la forma de expresión del significado[2] de la cabecera del artículo lexicográfico influye en el orden, por ejemplo la expresión para *acusar*: *Person V accuses person W in action X*. Esta expresión precede cada GP.

Otra información obligatoria en cada valencia sintáctica es la lista de todas las posibles formas de expresión de la valencia en los textos. El orden de opciones para una valencia dada es arbitraria, pero las opciones más frecuentes aparecen normalmente primero. Las opciones se expresan con símbolos de categorías gramaticales y palabras específicas.

A continuación presentamos una descripción para el verbo *acusar* aunque una descripción más amplia de este diccionario aparece en [Galicia et al, 98]. En esta descripción NP representa un grupo nominal e INF representa un verbo en infinitivo.

1 = V	2 = W	3 = X
1. NP Obligatoria	2. *a* NP Obligatoria	1. *de* NP 2. *de* INF

Posibles

C.1 + C.2	La policía *acusa* a Ana.
C.1 + C.2 + C.3.1	La policía *acusa* a Ana de robar.

Prohibidas:

C.1 + C.3.1	La policía *acusa* de robar.
C.3.1	*Acusa* de robo.

[2] Empleamos el inglés para la descripción de significado puesto que no existe un lenguaje semántico sin homonimia ni sinonimia, por lo que el inglés parece más conveniente que el mismo español para lectores hispanohablantes.

Para nuestro modelo, proponemos una nueva estructura de GP que llamamos *Patrones de manejo sintáctico avanzados* (ver Figura 2), que además de un formato modernizado para sistemas computacionales, incluye nuevos atributos para algunas características del Español (animidad en el objeto directo, repetición de valencias) y probabilidades para la realización y compatibilidad de valencias.

El trabajo manual ha sido la forma tradicional de compilar un diccionario de GP, por lo que su cobertura ha sido limitada. Para compilar este diccionario en lo que se refiere a información sintáctica, el método que proponemos para obtener los objetos de los verbos, sustantivos y adjetivos del español se basa en obtener las estadísticas de variantes del análisis sintáctico, las variantes son las combinaciones de palabras individuales con preposiciones. Si nos basamos solamente en categorías gramaticales, estas combinaciones serían las componentes de los denominados marcos de subcategorización pero específicos para cada palabra y estas palabras pueden ser verbos, adjetivos y sustantivos. La selección de este tipo de combinaciones o marcos de subcategorización específicos no es aleatoria. Esas combinaciones son fijas, en un buen grado, para cada palabra específica, así que sus estadísticas son más confiables que las de palabras arbitrarias. Los detalles de desarrollo se presentan en [Galicia *et al*, 2001].

El peso asignado a cada variante depende del número total de patrones y de valencias empatados, así como del tipo de patrones considerados, de las frecuencias de realización de las valencias y del número de homónimos en los patrones.

5 Red semántica

La proximidad semántica se refiere al conocimiento de contexto local. Cuando varias estructuras son igualmente posibles o el enlace de adjuntos (complementos no relacionados al significado de la palabra a la que se enlazan) es ambiguo, la proximidad semántica puede ayudar, es decir, los conceptos más cercanos relacionados a las palabras en los constituyentes posibles.

La idea detrás de la proximidad semántica es encontrar las trayectorias más cortas en una red semántica entre constituyentes obtenidos del módulo de reglas ponderadas. Aunque las redes semánticas son una aproximación a las habilidades humanas y por lo tanto son modelos simplificados, pueden usarse de una forma acorde a sus limitaciones.

Crear una red semántica es una tarea de labor intensa, y difícil de lograr aún a largo plazo. En este trabajo consideramos la red semántica que se está desarrollando a partir de la red FACTOTUM[3], mediante un método de traducción [Gelbukh, 98]. Para resolver la ambigüedad sintáctica, los enlaces de palabras o de grupos de palabras se realizan determinando el grado de proximidad semántica que tienen esas palabras o grupos de palabras.

La determinación de la proximidad semántica se basa en las características de la red semántica, que son: conceptos, relaciones, y trayectorias. Describimos la proximidad semántica como un valor cuantitativo, esta idea también ha sido empleada por [Sekine et al, 92; Rigau et al, 97]. Para determinarla no solamente consideramos la longitud por el número de enlaces sino también un peso asignado de acuerdo al tipo de relación. La trayectoria misma representa un valor cualitativo.

La proximidad entre un par de palabras es un valor que depende de la longitud y del tipo de relación. Para nosotros depende de las siguientes asignaciones:

- Un valor para cada tipo de relación
- Valores específicos para enlaces individuales
- Un valor mayor a relaciones implícitas.

La primera asignación contempla los valores mismos de las relaciones explícitas, es decir, su importancia. La segunda asignación pretende corregir el problema que se presenta conforme las relaciones están más cercanas al tope de la jerarquía, mientras más alejadas del tope, las palabras tienen más aspectos comunes.

La tercera asignación considera la problemática de las inferencias. Por ejemplo la relación *carro* ES_UN *objeto* y la relación implícita *objeto* TIENE_ SUBTIPO *libros*. De esta forma, la trayectoria es corta a pesar de que no hay muchos aspectos comunes. Para resolver este problema se asigna un peso mayor a una relación implícita que a una explícita. La precisión se obtiene junto con la segunda asignación que hace mayor la longitud de *carro* ES_UN *objeto* que de *Ford* ES_UN *carro*.

Desambiguación sintáctica. En el empleo de la red semántica para la desambiguación sintáctica realmente se está incorporando la componente semántica faltante en las reglas de CFG extendida. La estructura sintáctica en este módulo se toma de la salida producida en el módulo de reglas ponderadas. Algunas de las gramáticas más actuales, derivadas de las gramáticas generativas precisamente incorporan restricciones semánticas, como la HPSG que las considera en la entrada de cada lexema en el diccionario. Esto equivale a tener la red semántica interna de cada palabra con los vínculos a las posibles palabras con las que puede relacionarse en cualquier oración en el diccionario, lo cual implica una labor intensa.

En nuestro modelo, esas restricciones semánticas se buscan en la red y se definen a través de la proximidad semántica, que involucra la distancia menor entre pares de palabras y su valor asignado. La evaluación de la proximidad no sólo está relacionada con estos valores obtenidos de la red misma, como se mostró anteriormente, sino que es necesario considerar además el tipo sintáctico de la relación. No todas las trayectorias son aceptables en un contexto específico. En algunos casos se tendrá que buscar la trayectoria con las relaciones que sean más adecuadas al contexto sintáctico de la oración. Por ejemplo, en la frase *Veo un gato con un telescopio* aparece la frase preposicional *con un telescopio*, la relación más cercana será USO y una relación más cercana tipo ES_UN no será la más adecuada para ese contexto (Figura 3).

[3] *FACTOTUM® SemN et*, es una red semántica compilada por la empresa MICRA, INC. New Jersey, USA.

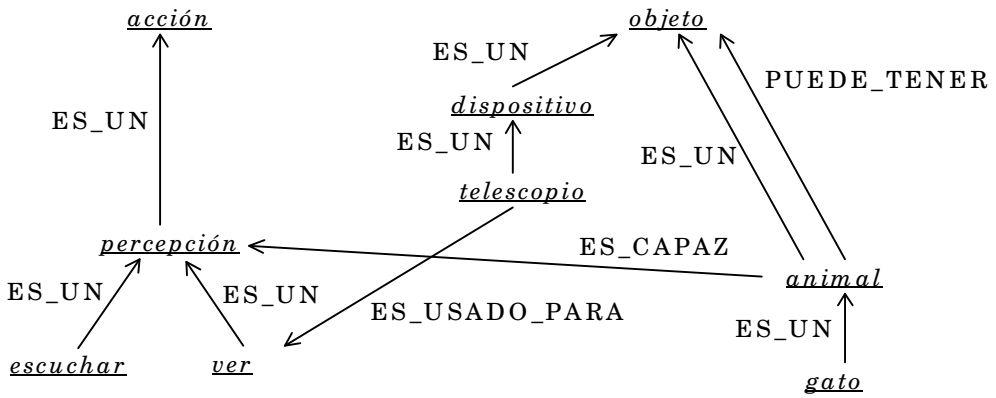

Figura 3. Fragmento de la red semántica para la frase *Veo un gato con un telescopio.*

Así que la tarea de desambiguación está muy relacionada con el método para encontrar las trayectorias aceptables mínimas y de contexto sintáctico. En una red semántica existe un número infinito de trayectorias conectando dos palabras. Los aspectos matemáticos de solución y de implementación computacional se describen en [Gelbukh, 98].

6 Módulo de votación

Para desambiguar estructuras sintácticas un módulo de votación emplea los pesos asignados en cada módulo, vota por el máximo valor sumado de las variantes (ver Figura 4). El resultado es una lista clasificada de las variantes sintácticas.

Para poder hacer la votación entonces, nuestro modelo requiere una evaluación cuantitativa para ordenar las variantes construidas por cada módulo, y una forma que las haga compatibles para su evaluación. La compatibilidad se logra mediante un formato común que es la estructura de dependencias. Se transforman las estructuras de constituyentes a una estructura de dependencias, para hacer posible la comparación de sus valores con los valores de las estructuras del módulo de GP.

A continuación presentamos un ejemplo, para la frase *El productor trasladó la filmación de los estudios al estadio universitario*, indicando solamente las ideas del modelo. En este ejemplo consideramos lo siguiente:

- Patrones de manejo:

4.34896 *trasladar*,dobj_suj,obj:a,obj:de, x:?
0.43697 *trasladar*, obj:a, x:?
1.13758 *filmación*
3.29976 *estadio*

donde los números de la primera columna representan los valores obtenidos del método de compilación de información sintáctica para los patrones de manejo. La marca "x:?" representa una valencia repetida mediante clíticos, "dobj_suj" indica el grupo nominal que puede ser sujeto u objeto directo, "obj" indica los complementos preposicionales y enseguida la preposición específica.

Así que considerando los patrones de *trasladar*, *estadio* y *filmación* se favorecen las variantes con la estructura de: trasladar algo a algún lugar desde otro lugar.

- Con el modelo de reglas ponderadas obtenemos 8 variantes con el mismo peso.
- Con la proximidad semántica, encontramos las siguientes relaciones:

filmación → *director*
→ subtipo de *espectáculo*
trasladar → con referencia a una dirección o a un lugar
→ con relación a traslación de un objeto
→ subtipo *trayectoria*
Estudio cinematográfico → *lugar*
estadio → como subtipo de *espectáculo*
filmación → *cine* → *director*

Así que únicamente la relación entre *trasladar* y *estudio* como *lugar* puede

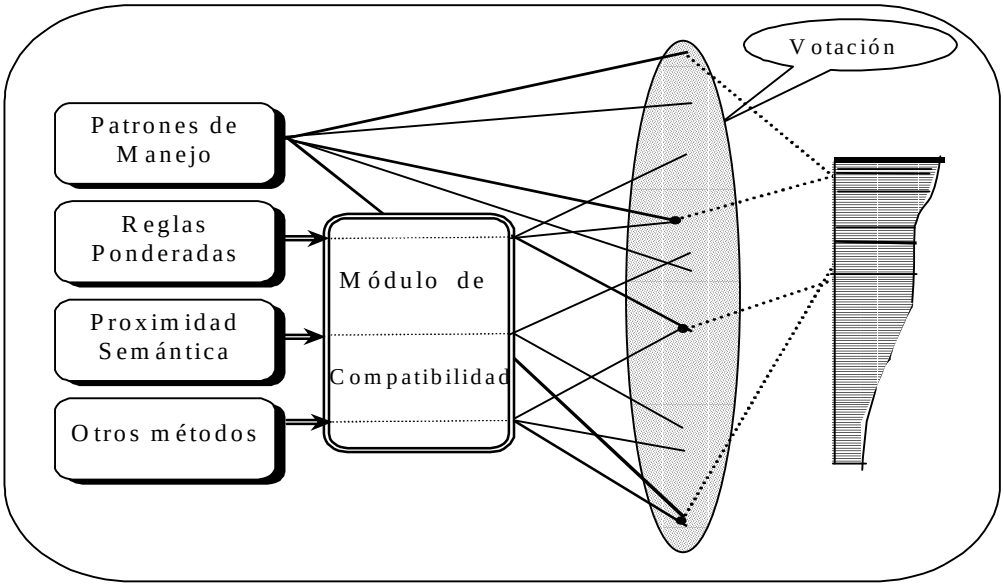

Figura 4. Modelo de análisis sintáctico y desambiguación

considerarse, favoreciendo la estructura: trasladar desde lugar (*trasladar de los estudios*).

En este ejemplo el método de patrones de manejo es el que más contribuye para reconocer las variantes correctas, seguido por la información del modelo de proximidad semántica.

Para valorar la precisión de los métodos desarrollados para desambiguación se han considerado diversos elementos de evaluación. Por ejemplo, enlaces entre palabras de contenido [Yuret, 98] comparados contra un conjunto de oraciones analizadas manualmente, o el número de constituyentes obtenidos [Collins, 99] comparado contra un corpus marcado sintácticamente.

En nuestro caso consideramos la aparición de las variantes más probables de ser las correctas en un rango en el tope de la clasificación, utilizando un conjunto de oraciones analizadas manualmente.

En experimentos realizados, los resultados obtenidos en la aplicación de nuestro método de compilación de patrones de manejo sintáctico al corpus LEXESP fueron usados para analizar sintácticamente 100 oraciones y las estructuras verdaderas se encontraron clasificadas en el rango tope del 35%.

7 Conclusiones

Proponemos un esquema de análisis sintáctico que considera la inclusión de tres fuentes de conocimiento: léxica, sintáctica y semántica. Sólo con la participación de estos conocimientos es posible diferenciar las variantes sintácticas correctas de entre las múltiples variantes sintácticas generadas en el proceso de análisis sintáctico.

Los mecanismos que proponemos para la resolución de ambigüedad estructural son los Patrones de manejo sintáctico que reflejan conocimiento léxico y sintáctico, las Reglas ponderadas que reflejan conocimiento sintáctico, y la Red semántica que refleja conocimiento semántico de cercanía de sentido entre grupos sintácticos.

Mostramos cómo contribuyen estos conocimientos para revelar la estructura sintáctica correcta de las oraciones analizadas.

Experimentos realizados, aplicando nuestro método de compilación de patrones de manejo sintáctico a un corpus de 100 oraciones muestran las estructuras verdaderas en el rango del 35%.

Nuestro objetivo es el análisis de textos sin restricciones por lo que se desarrollaron métodos automáticos de compilación de los recursos requeridos. Las herramientas desarrolladas tienen la gran ventaja de no requerir ninguna desambiguación de marcas

morfológicas ni ningún tipo de marcas sintácticas.

Referencias

[Charniak, 97] Charniak, E. *Statistical parsing with a context-free grammar and word statistics,* Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI MIT Press, Menlo Park, 1997. http:// www.cs.brown.edu/people/ec/home.html.

[Civit & Castellón, 98] Civit, M e I. Castellón. *Gramesp: Una gramática de corpus para el español*. Revista de AESLA, La Rioja, España, 1998.

[Collins, 96] Collins, M. *A new Statistical Parser Based on Bigram Lexical Dependencies*. In Proceedings 34th Annual Meeting of ACL, pp.184-191. 1996.

[Collins, 99] Collins, M. *Head-driven Statistical Models for Natural language parsing*. Ph.D. thesis. University of Pennsylvania. 1999. http://xxx.lanl.gov/ find/ cmp-lg.

[Galicia *et al*, 98] Galicia Haro, S., I. Bolshakov, A. Gelbukh. *Diccionario de patrones de manejo sintáctico para análisis de textos en español*. Revista Procesamiento de lenguaje natural, No. 23, SEPLN, España, pp. 171-176. Septiembre de 1998.

[Galicia *et al*, 2001] Galicia Haro, S., I. Bolshakov, A. Gelbukh. *Obtención semiautomática de patrones de manejo para el lenguaje español*. Segundo Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje, por aparecer.

[Gelbukh, 98] Gelbukh, A. F. *Lexical, syntactic, and referencial disambiguation using a semantic network dictionary*. Technical report. CIC, IPN, 1998

[Magerman, 95] Magerman, D. M. *Statistical decision-Tree Models for Parsing*. In Proceedings 33rd Annual Meeting of ACL. June 26-30 Cambridge, Massachusetts, USA, pp. 276-283, 1995. http://xxx.lanl.gov/ps/cmp-lg/9504030

[Mel'cuk, 88] Mel'cuk, I. A. *Dependency Syntax: Theory and Practice*. State University of New York Press. Albany

[Rigau *et al*, 97] Rigau, G., Atserias, J. and Agirre, E. *Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 48-55, 1997 http://xxx.lanl.gov/ps/cmp-lg/9704007

[Sekine *et al*, 92] Sekine, S., Carroll, J. J., Ananiadou, S. and Tsujii, J. *Automatic Learning for Semantic Collocation*. In Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, pp. 104-110, 1992.

[Steele, 90] Steele, J. *Meaning – Text Theory. Linguistics, Lexicography, and Implications.* J. Steele (ed.). University of Ottawa press.

[Yuret, 98] Yuret, D. *Discovery of Linguistic Relations Using Lexical Attraction*. Ph. D. thesis. Massachusetts Institute of Technology. 1998. http:// xxx.lanl. gov/find/cmp-lg/9805009