Diccionario de patrones de manejo sintáctico para análisis de textos en español

S. N. Galicia Haro, A. F. Gelbukh, I. A. Bolshakov
Laboratorio de Procesamiento de Lenguaje Natural,
Centro de Investigación en Computación, IPN
Av. Juan de Dios Batiz s/n esquina Miguel Othón de Mendizabal,
Unidad Profesional Adolfo López Mateos. 07738 México D.F.
{sofia, gelbukh}@pollux.cic.ipn.mx, iabolsh@aha.ru

Resumen

Se presenta el desarrollo de un diccionario de patrones de manejo sintáctico para el español que servirá para el análisis sintáctico de textos arbitrarios en español y para enseñanza del idioma a extranjeros.

Se presenta la primera tarea de descripción de un conjunto de patrones de manejo sintáctico con cerca de 600 verbos y 150 adjetivos del español. Este conjunto se tomó de varias fuentes y cubre algunos verbos de la vida cotidiana y de la esfera política.

Los patrones de manejo sintáctico son una colección completa de descripciones de todos los posibles "objetos" de una palabra especifica: verbos, adjetivos y sustantivos. Estos patrones también contienen, en la descripción, los órdenes permitidos y prohibidos de estas diferentes palabras manejadas.

En este método de descripción basado en las gramáticas con dependencia, las palabras subordinadas (o manejadas) fuertemente a otras palabras como sustantivos, verbos y adjetivos, se estudian sin considerar su orden en la oración. Lo que permite aplicarlo a lenguajes naturales con restricciones en el orden de palabras diferentes del inglés, como el español.

Se presentan las estadísticas de valencias sintácticas y preposiciones empleadas en estos Patrones de Manejo Sintáctico y se comparan algunos valores obtenidos contra estudios realizados.

Palabras clave: diccionario, patrones de manejo sintáctico, valencias sintácticas, análisis sintáctico, gramáticas con dependencia

1. Introducción

La descripción de estructuras sintácticas mediante diversas gramáticas se ha planteado desde los inicios del procesamiento del lenguaje natural principalmente basadas en las Gramáticas Independientes del Contexto (CFG en inglés) [Allen 94] y con sus diferentes extensiones para considerar las características sensitivas al contexto.

Sin embargo no existe un método que haya sido utilizado satisfactoriamente para analizar completamente cualquier lenguaje natural sin restricciones. Aunque con el inglés se han obtenido algunos buenos resultados, este éxito se debe principalmente a que en este lenguaje, el ordenamiento de los miembros dentro de la oración es estricto, generalmente.

Para la lengua española, el orden de palabras es mucho más libre. De hecho, si una oración en español contiene N grupos de palabras subordinadas al predicado de una oración, y si cada grupo puede ocupar un lugar arbitrario dentro de la oración, el número de reglas independientes del contexto para el nivel superior de la oración puede calcularse como factorial de N. Por ejemplo, en una oración muy simple existe una gran variación de construcción: *Juan vino a mi casa, A mi casa vino Juan, Vino Juan a mi casa, A mi casa Juan vino, Juan a mi casa vino, Vino a mi casa Juan.*

El método alternativo de descripción de estructuras sintácticas que se propone se basa en las gramáticas con dependencia (DG en inglés). Las DG suponen que cada palabra de la oración está subordinada a otra de tal forma que se crea un árbol de dependencias

sintácticas de la oración. La única excepción es la raíz del árbol, en cuyo lugar se sitúa normalmente un verbo predicativo en forma personal. De esta manera, se separa la información de la subordinación y de la posición de la palabra subordinada, respecto a la que subordina.

En las DG, las palabras subordinadas (o manejadas) fuertemente a los verbos, adjetivos y sustantivos se estudian sin considerar su orden en la oración. Los patrones de manejo sintáctico son una colección completa de descripciones de todos los posibles "objetos" de una palabra específica. Se presenta un ejemplo de patrones de manejo sintáctico siguiendo el método de descripción del modelo Significado ⇔ Texto (*Meaning ⇔ Text*) realizado por Igor Mel'čuk [Steele 90] y un ejemplo de desambiguación sintáctica basada en estos patrones.

2. *Información en los patrones de manejo sintáctico*

En los diccionarios comunes, y por supuesto del español, la única marca establecida para valencias de los verbos es la de transitividad o intransitividad. Que el verbo sea transitivo implica que la acción se transfiere del sujeto a un objeto explícito y por lo tanto el verbo contiene por lo menos dos valencias sintácticas: sujeto y objeto. Los verbos intransitivos no tienen objeto directo, aunque algunas veces el mismo verbo puede utilizarse con otro significado como transitivo, por ejemplo: *vive muy bien* y *vive su vida*. En los diccionarios comunes no se incluye la información de las otras valencias de los verbos, ni la de las valencias de los sustantivos y adjetivos.

En cambio, en el diccionario de patrones de manejo sintáctico se describen *todas* las valencias de los verbos de acuerdo a su significado, así como las valencias de los sustantivos y adjetivos [BBI 86]. En el español existen verbos con 0 valencias sintácticas y hasta 5, en general; sin embargo, la mayoría de los verbos en español caen en el rango de 1 a 3 valencias. Ejemplos de oraciones cuyos verbos tienen diferentes números de valencias:

Llueve.
Juan vive mal.
Juan mira las montañas.
Juan acuerda el proyecto con su jefe.
Juan compra un vestido en la tienda en 5 pesetas.
Juan renta un departamento a la señora María por un año en 500 pesetas.

Otra diferencia con los diccionarios comunes es que en éstos no se marcan las preposiciones de enlace, que van ante los complementos de verbos, adjetivos y de algunos sustantivos [DRAE 95]. Este conocimiento no se puede establecer mediante un algoritmo porque no es un conocimiento lógico, es decir, no hay una razón para que el sustantivo *querella* emplee la preposición *contra*, por ejemplo: *querella contra Juan* o para que el sustantivo *solidaridad* use la preposición *con: solidaridad con Juan* para indicar el receptor. Por lo que la única manera de establecerlo es mediante un diccionario.

Los complementos de los verbos exigen el empleo de una determinada preposición. Ejemplos: *me arrepiento de mis acciones, lo expresó con ademanes, insisto en pagar*. Esto ocurre también con sustantivos y adjetivos que exigen el empleo de una determinada preposición. Ejemplos: *intolerante con sus amigos, esencial en el proyecto, inferior a su compañero*. En cuanto al objeto directo, normalmente se construye sin preposición salvo cuando designa seres humanos o animados que podrían aparecer en la posición de sujeto. En estos patrones de manejo sintáctico se hará notar la posibilidad de aparición de la preposición si así lo requiere el verbo, por ejemplo para *ver*: *veo al perro de María* y *veo el perro, la casa* y *el patio* . En este caso, se utilizará la notación: *ver (a)* para indicar su posibilidad de aparición.

En estos patrones de manejo sintáctico, se describen las preposiciones específicas que enlazan a los verbos, adjetivos y sustantivos con sus complementos, lo cuál permite eliminar ambigüedad sintáctica.

3. *Estructura del diccionario*

Las entradas del diccionario de patrones de manejo sintáctico constan de cuatro secciones:

La palabra encabezado.

La explicación semántica.

Las valencias sintácticas de la palabra encabezado.

Los órdenes posibles e imposibles.

La palabra encabezado corresponde al verbo, sustantivo o adjetivo considerado, con un significado específico. Para diferenciar los patrones de manejo sintáctico de palabras homónimas, se da una numeración, por ejemplo: *alternar*$_1$ ('tener trato con otras personas') y *alternar*$_2$ ('hacer dos o más acciones una tras otra y repetidamente'). Para diferenciarlos, la numeración es totalmente arbitraria pero debe existir al menos un elemento diferente en el patrón de manejo sintáctico respecto de los otros.

La explicación semántica de la situación relacionada a cada palabra específica. En nuestra colección se ha optado por una simplificación del método de descripción del modelo Significado ⇔ Texto, la explicación semántica se reemplaza por una oración simple en inglés.

Las valencias sintácticas de la palabra encabezado. Los principios de ordenamiento de las valencias de los verbos son exactamente como en la teoría Significado ⇔ Texto, primero aparece el sujeto, después el objeto directo y luego los objetos indirectos comenzando por los más importantes para la situación normal relacionada con la palabra, aunque claro está, que para algunas personas este orden puede resultar algo subjetivo.

Los órdenes posibles e imposibles constituyen la última sección y muestran los órdenes y combinaciones posibles e imposibles mediante números de valencias y las variantes de realización. En esta sección la palabra encabezado está representada por el número cero.

A continuación se presenta un ejemplo de los patrones de manejo sintáctico, para el verbo *solicitar*. Las preposiciones se marcan en tipo itálico, la tilde se utiliza para colocar la palabra encabezado.

solicitar

X asks something Y from Z

Número	Patrón de manejo	Ejemplo
X = 1; who asks?		
1.1	S (an)	Juan / el gobierno ~
Y = 2; what?		
2.1	S (na)	~ una prórroga / un préstamo
2.2	*que* C	que este libro se le dé
Z = 3; from whom?		
3.1	*a* S (an)	a la secretaria
3.2	*con* S (an)	con el secretario
3.3	*de* S (an)	de usted
3.4	*en* S (na)	en urgencias
POSIBLE:		
(1) 0 2 3	(El partido) solicita una prórroga al gobierno	
(1) 0 2	(Ella) solicita un préstamo	
IMPOSIBLE:		
(1) 0	*(El partido) solicita	
(1) 0 3	*(El partido) solicita al gobierno.	

donde:

S - sustantivo o pronombre personal

que C - claúsula subordinada relacionada a la principal a través de *que* (... *que este libro se dé al muchacho*)

(an) - animado (solamente para sustantivos), corresponden a criaturas vivientes, incluyendo al ser humano, grupos de humanos, organizaciones, etc.,

(na) - inanimado (solamente para sustantivos), como argumento, acción y lugar.

Existen otras abreviaturas que no se utilizaron en este ejemplo, como:

V - verbo

Adj - adjetivo

Adv - adverbio

Pp - pronombre personal

Q - claúsula subordinada que tiene forma de interrogación. Por ejemplo, para el verbo *decir*: "*Dijo: ¿A quién se dió este libro?*"

(inf) - forma infinitiva (solamente verbos)

(tm) - intervalo de tiempo (solamente para sustantivos)

(mn) - de manera (solamente para adverbios)

(pc) - de lugar (solamente para adverbios)

(nom) - caso nominativo (solamente para pronombres personales)

(acc) - caso acusativo (solamente para pronombres personales)

(dat) - caso dativo (solamente para pronombres personales)

(inc) - caso inclusivo (solamente para pronombres personales)

Este caso inclusivo lo introducimos como una designación de las formas contraídas *conmigo, contigo, consigo.*

El comienzo de este diccionario, con cerca de 600 verbos y 150 adjetivos del español, cubre algunos verbos y adjetivos de la vida cotidiana y de la esfera política. Estas palabras se tomaron de dos fuentes: diversas gramáticas y libros de enseñanza del español para extranjeros, y de periódicos mexicanos actuales. Esta colección es una base para obtener las peculiaridades de los patrones de manejo sintáctico de verbos, adjetivos y sustantivos del español, posiblemente en su totalidad.

4. *Aplicación para resolución de ambigüedad sintáctica*

El conjunto de estos patrones de manejo sintáctico formará un diccionario combinatorio, que servirá para realizar el análisis sintáctico. Por ejemplo: si se tiene un diccionario combinatorio de la siguiente forma:

solicitar	∅ [algo]
	a, con, de [alguien]
plática	*sobre* [algo, alguien]

entonces, se pueden distinguir fácilmente las estructuras de las dos frases siguientes usando este diccionario para verificar o marcar las relaciones:

Solicitó una plática sobre N. Chomsky.

Solicitó una plática a N. Chomsky.

Ambas oraciones tienen los mismos constituyentes: el mismo verbo, el mismo objeto directo y una frase preposicional, la diferencia radica en la preposición específica en cada oración. En la primera oración la frase preposicional *sobre N. Chomsky* está unida sintácticamente al sustantivo *plática* y en la segunda oración, la frase preposicional *a N. Chomsky* está unida al verbo *solicitar* porque en el diccionario de patrones de manejo sintáctico el uso de la preposición *a* esta asociada con la tercera valencia del verbo *solicitar.*

En algunos casos, el patrón de manejo sintáctico del verbo y del sustantivo pueden coincidir en el uso de una preposición específica, como en el caso *solicitar (un permiso) con el director, solidaridad con el director.* En este caso se utilizarían pesos asignados a enlaces, es decir, el uso más frecuente que se utiliza en el lenguaje, por ejemplo: usos obtenidos a partir de un corpus. Lo que permitiría reconocer que para la palabra *solidaridad* la preposición *con* tiene un uso mayor en frecuencia que con *solicitar.*

En el ejemplo de este diccionario, intencionalmente se usó una estructura muy simple sólo para ilustrar la idea. Un diccionario real tiene mucha más información acerca de cada palabra. Para realizar el análisis sintáctico mediante esta aproximación, el diccionario de patrones de manejo sintáctico para el español sería la principal fuente de datos para un analizador guiado por estos patrones. Este diccionario listaría las valencias de las palabras, los medios para expresarlas, es decir, las preposiciones correspondientes, las propiedades de las palabras que pueden llenar estas valencias, y la información adicional en la interpretación, o significado de estas valencias. También tendría la información acerca de la posibilidad o imposibilidad de aparición de ciertas combinaciones de valencias en la misma oración.

5. *Algunas estadísticas*

Las estadísticas que presentamos a continuación servirán de base para la siguiente fase de la compilación de este diccionario para el español que incluirá la obtención semiautomática de estos patrones a partir de un corpus.

El número de valencias que presentan los 612 verbos y 142 adjetivos se muestran en la siguiente tabla, lo que confirma que la mayoría de los verbos del español tienen entre 2 y 3 valencias. En los adjetivos la primera valencia corresponde al sustantivo y por lo tanto no existen patrones de adjetivos que solamente contengan una valencia.

Número de valencias	Cantidad de verbos	Cantidad de adjetivos
1	3	–
2	363	111
3	210	31
4	34	–
5	1	–
6	1	–

El número de preposiciones que presentan los 612 verbos se muestran en la siguiente tabla. Tanto para verbos de dos valencias como de tres valencias, alrededor del 60% determinan la segunda valencia con una sola preposición, usualmente *a*. En cambio determinan la misma valencia con dos preposiciones diferentes el 20% de los verbos con dos valencias y el 6% de los verbos con tres valencias. Lo que implica que aunque un porcentaje alto de frases que contengan los verbos aquí considerados se podrían desambiguar sintácticamente con métodos de enlazado de frases preposicionales de una sola preposición, para cubrir el total se requieren métodos que contemplen el enlace de varias frases preposicionales.

Número De Prepo-siciones	Verbos Con 2 valencias (Val. Y)	Verbos con 3 valencias (Val. Y)	(Val. Z)
0	44	70	15
1	227	126	154
2	72	12	35
3	17	2	6
4	3	–	–

El número de preposiciones que presentan los 142 adjetivos se muestran en la siguiente tabla. Para adjetivos de dos valencias el 54% determinan la segunda valencia con una sola preposición y el 29% con dos preposiciones. En cambio de los adjetivos con tres valencias, el 64.5% determinan la segunda valencia con una preposición y el 10% con dos preposiciones diferentes. Lo que representa porcentajes similares que en los verbos. Así que para los adjetivos se debe hacer la misma consideración de métodos que contemplen el enlace de varias frases preposicionales para desambiguación.

Número de Preposi-ciones	Adjetivos con 2 valencias (Val. Y)	Adjetivos con 3 valencias (Val. Y)	(Val. Z)
0	–	–	–
1	60	20	29
2	32	3	2
3	16	7	–
4	2	1	–

Derivada de estas colecciones de patrones de manejo sintáctico, se obtuvo la siguiente tabla que presenta el orden de las preposiciones simples referido a su frecuencia de aparición. Mientras que la preposición *a* fue la que más apareció en patrones de verbos, la preposición *de* fue la que más se presentó en los patrones de adjetivos.

No.	Posición en patrones de verbos	adjetivos
1	a	de
2	de	en
3	en	a
4	con	para
5	por	con
6	contra	por
7	para	contra
8	sobre	–

Por otro lado, tenemos las preposiciones más frecuentes, empleadas en el léxico mexicano, que fueron reportadas por el Colegio de México [Ham 79]. A continuación se presentan con la posición obtenida dentro de las 100 palabras más frecuentes.

Posi-ción	Preposi-ción
3	De
6	En
7	A
12	Por
13	Con
17	Para

Posi-ción	Preposi-ción
55	Sin
56	Hasta
66	Sobre
71	Entre
84	Desde

En la siguiente tabla relacionamos las preposiciones por su orden de aparición en estos trabajos:

Preposi-ción	Posición en CM	Posición en verbos	Posición en adjetivos
De	1	2	1
En	2	3	2
A	3	1	3
Por	4	5	6

Con	5	4	4
Para	6	7	5
Contra	–	6	7
Sobre	9	8	–

Para medir la aproximación de estos ordenamientos obtuvimos los coeficientes de rango de correlación basados en el coeficiente de correlación de rango de Kendall.

Coeficiente de rango de correlación hacia el ordenamiento en CM:

En patrones de manejo sintáctico para verbos: 0.64

En patrones de manejo sintáctico para adjetivos: 0.86

Estos coeficientes de rango de correlación indican la medición de desarreglo que existe en relación al ordenamiento encontrado por el Colegio de México. En el caso de los patrones de manejo sintáctico para adjetivos, la interdependencia es mayor y efectivamente las posiciones son muy parecidas a excepción de la preposición *por* que pasa de la cuarta a la sexta posición. En cambio para el caso de los patrones de manejo sintáctico para verbos, la interdependencia no es tan cercana a uno, la preposición *a* aparece con mayor frecuencia porque se relaciona con los complementos directos e indirectos. Probablemente, al tener un mayor número de patrones de manejo sintáctico para el español la interdependencia se incremente.

6. Conclusiones

Para el desarrollo de cualquier sistema de extracción de información a partir de textos arbitrarios en español, tales como libros, periódicos u otras fuentes textuales sin preparación previa, una parte importante del trabajo será desarrollar un analizador sintáctico adecuado al español.

Como ya mencionamos, aún para el inglés, ha habido problemas al querer analizar textos arbitrarios mediante distintas aproximaciones principalmente por la gran cantidad de reglas de sustitución que se requieren y de árboles sintácticos que mediante ellas se producen. En los patrones de manejo sintáctico, se describen las preposiciones específicas que enlazan a los verbos, adjetivos y sustantivos con sus complementos, lo cuál permite eliminar ambigüedad sintáctica. Por lo que consideramos que la alternativa de descripción que se propone es más adecuada al español donde el orden de palabras es más libre y el uso de preposiciones es más amplio.

Esta alternativa mediante patrones de manejo sintáctico, no ha sido empleada intensivamente por otros investigadores, debido a que la composición del diccionario necesita muchísimo esfuerzo en si mismo. La siguiente fase de la compilación de este diccionario para el español incluye la obtención semiautomática de estos patrones a partir de un corpus. actualmente se experimenta con un algoritmo que se describe en [Gelboukh et al, 98].

Bibliografía

[Allen 94] Allen, J. F.Natural language understanding. Benjamin Cummings. 1994.

[BBI 86] Benson, M., E. Benson, and R. Ilson. The BBI combinatory dictionary of English. John Benjamins Publishing Co., 1986.

[DRAE 95] Real Academia Española. Diccionario de la Lengua Española. Edición vigésima primera, en CD-ROM de ESPASA CALPE. 1995.

[Gelbouk et al, 98] Gelbukh, A.F, Bolshakov, I.A., Galicia Haro, S.N. Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts. *CONALD-98: Conference on Automatic Learning and Discovery*, Carnegie Mellon University, Pittsburgh, PA, USA. June, 1998.

[Ham 79] Ham Chande, Roberto. Del 1 al 100 en Lexicografía. En: *Investigaciones Lingüísticas en Lexicografía*. Jornadas 89. México: El Colegio de México, 1979, pag. 41-83.

[Steele 90] Steele, J., editor. Meaning – Text Theory. Linguistics, Lexicography, and Implications. University of Ottawa Press, 1990.